Doctoral theses at NTNU, 2022:354

Wouter Koch

Improving the citizen science data corpus for science and management

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Natural Sciences
Department of Biology

NTNU
Norwegian University of
Science and Technology

Wouter Koch

Improving the citizen science data corpus for science and management

Thesis for the Degree of Philosophiae Doctor

Trondheim, November 2022

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Natural Sciences
Department of Biology

ISBN 978-82-326-6083-4 (printed ver.)
ISBN 978-82-326-6669-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:354

Printed by NTNU Grafisk senter

Summary

Citizen science, in which amateur volunteers report their observations, is becoming an increasingly important source of biodiversity data. To understand and properly manage natural resources, we need large amounts of observational data, and contributions from citizens are crucial in obtaining that. All observational data, but especially from citizen scientists, come with a number of challenges that we need to be aware of, and where possible, address. These include the need for open access to existing data, and the fact that some species are more popular and/or easier to recognize so that they are reported more, especially when it comes to pictures which are important for image recognition models. Additionally, correctly identifying species requires expert knowledge which citizens do not always have access to and which is becoming more rare in general, so such knowledge needs to be stored in a systematic way.

In this thesis, the aim is to investigate how widespread these issues are in data from citizen science, and what we can do to minimize them at the data collection stage.
To do so, I have:

- reviewed how researchers that use open data also openly share the data they add
- evaluated how pictures taken by citizen scientists help improve AI image recognition, and how this relates to how popular the different species groups are
- investigated how AI image recognition and the number of pictures that are available relate to both the quality of pictures and how easy the species are to recognize
- proposed a new data format for identification keys, so that experts can store their knowledge and citizens (among others) can better identify species

A lot of effort goes into methods for how to deal with issues in citizen science data in terms of coverage, quality and biases. Meanwhile, we should not be complacent, and keep aiming for better collection methods that minimize these issues to begin with, and look for other ways to improve the biodiversity data needed for research and management.

Sammendrag

Folkeforskning, hvor frivillige amatører rapporterer sine observasjoner, er en stadig viktigere kilde for biodiversitetsdata. For å kunne forstå og forvalte naturressurser trenger vi store mengder observasjonsdata, og bidrag fra folkeforskere er avgjørende for å få tak i det. Alle observasjonsdata, men særlig fra folkeforskning, medfører en del utfordringer som man må være oppmerksom på, og hvor mulig gjøre noe med. Disse er blant annet behovet for åpen tilgang til eksisterende data, og faktumet at noen arter er mer populære og/eller lettere å gjenkjenne slik at de blir rapportert oftere, særlig når det gjelder bilder, som er viktige for automatisk gjenkjenning av arter. I tillegg krever riktig artsbestemmelse ekspertkunnskap som folkeforskere ikke alltid har tilgang til og som i utgangspunktet blir mer sjelden, så slik kunnskap må lagres på en systematisk måte.

I denne avhandlingen er målet å undersøke hvor utbredt disse utfordringer er i dataene vi får inn fra folkeforskning, og hva vi kan gjøre for å minimalisere de i datainnsamlingsfasen.
For å gjøre det har jeg:

- kartlagt hvordan forskere som bruker åpne data også gjør dataene de legger til åpent tilgjengelig
- evaluert hvordan bilder tatt av folkeforskere hjelper i å forbedre bildegjenkjenning ved hjelp av kunstig intelligens, og hvordan dette forholder seg til hvor populære de ulike artsgruppene er
- undersøkt hvordan bildegjenkjenning ved hjelp av kunstig intelligens, samt hvor mange bilder som er tilgjengelige forholder seg til bildekvalitet og hvor lett det er å gjenkjenne artene
- foreslått et nytt dataformat for bestemmelsesnøkler, slik at eksperter kan lagre sin kunnskap og (blant annet) folkeforskere kan bli bedre til å bestemme arter

Mye energi brukes på metoder for hvordan vi kan håndtere utfordringer i folkeforskningsdata når det gjelder dekning, kvalitet og skjevhet i datagrunnlaget. Samtidig bør vi ikke ta disse problemene for gitt, og fortsette å prøve å få til bedre innsamlingsmetoder som minimaliserer disse utfordringer i utgangspunktet, samt se etter andre måter for å forbedre biodiveritetsdataene som vi trenger for forskning og forvaltning.

Contents

Acknowledgements

One of the beautiful things about science is that it allows us to bumble along, getting it wrong time after time, and feel perfectly fine as long as we learn something each time.

Martin A. Schwartz[1]

To me, this is the most important part of my thesis by far, the part where I get to thank the many people that have been there for me, each in their own way. It has been an unforeseen privilege to get to know many wonderful new people, and it is bittersweet to end this period.

I started my PhD years after finishing my master's, and it was far from an obvious path to pursue. Leaving my comfort zone and embracing my productive stupidity[1] has been one of the best choices I ever made. Without the encouragement from Anders, and the knowledge that he would be supervising me I would probably never even have attempted it at all. The support from Artsdatabanken and the Norwegian Research Council opened doors I didn't know existed, and I am still amazed and slightly embarrassed by how stupendously fortunate I am to have been given this opportunity.

My supervisors Anders, Erlend, Bob, and Vidar have been very patient with me, which was needed as I never seem to quite know what I want except that my plans need to change drastically, daresay significantly[2]. Thank you all for your support. My colleagues at Artsdatabanken have been there for me along the way, for cafecito breaks and a chat when I needed it, but also by giving me the space and quiet when I needed that. I look forward to making my comeback there. At NTNU, I got to be part of a diverse group of students from different disciplines; the Transforming Citizen Science for Biodiversity group. To Ben, Caitlin, Jan, Jorge, Kwaku, and Philip; thanks for the cooperation and the fun times, I hope there will be more to come of both.

While the pandemic has resulted in more home office days than are good for a person, I have been so lucky to be located at various great institutions. Artsdatabanken, the University Museum, IBI, CBD, NINA, Naturalis, LIACS were all places I felt right at home. I will really miss being at the CBD as often, it is a privilege to hang out with such smart, kind people from all around the world during coffee breaks, lunch breaks, PhD meetings, retreats, and social events after work. Thanks so much for having me.

It usually takes me about 5 years to come out of my shell, which is unfortunate as a PhD only lasts 4 years. Being welcomed by many friendly, inclusive people

has helped me a lot. A special thanks in this respect to my roomies Lise at LIACS and Myranda at the CBD.

I have had little to complain about, but when I did, I did so in the best company possible. Thanks Sam, Stefan, Mari, Archana, Caitlin, Ingeborg, Philip and Stacy. Whining and dining with you kept me sane.

Laurens, you have helped me as both a friend, a collaborator, and an example in making the bold choice of working with that interdisciplinary niche interest we happen to share. My time at Naturalis in Leiden was an opportunity I had dreamt of in many respects for me and my family. Thank you for your help in making it happen, the lunch strolls in Leiden, and the great time all of us had while we were there.

Rim and Rienk, it's a shame we are all so far apart these days, but your end-to-end encrypted nonsense has distracted me from my work constantly. Thanks.

Aaron Swartz and Alexandra Elbakyan, thanks for your important work and practical aid.

Finally, some words of gratitude to my family, truly the foundation beneath it all. Mom and dad, thanks for the support all these years, which has really been beyond what anyone can reasonably expect, even from family. Petra, zusje, I'm proud of you.

None have been more supportive than my girlfriend Ragnhild and our kids Marius, Jonas, and Nora. I know it has not been easy living with me being a student, with everything that -apparently- comes with that. I would never in a million years have even tried anything like this if it had not been for your backing. You are the best ♡

List of articles

This thesis contains the following articles:

1. Mandeville, C. P., **Koch, W.**, Nilsen, E. B. & Finstad, A. G. Open Data Practices among Users of Primary Biodiversity Data. *BioScience* **71**, 1128–1147. https://doi.org/10.1093/biosci/biab072 (Aug. 2021).

2. **Koch, W.**, Hogeweg, L., Nilsen, E. B. & Finstad, A. G. Maximizing citizen scientists' contribution to automated species recognition. *Scientific Reports* **12**. https://doi.org/10.1038/s41598-022-11257-x (May 2022).

3. **Koch, W.**, Hogeweg, L., Nilsen, E. B., O'Hara, R. B. & Finstad, A. G. Recognizability bias in citizen science photographs. https://doi.org/10.1101/2022.06.25.497604 (Jun. 2022). Preprint.

4. **Koch, W.**, Elven, H. & Finstad, A. G. Clavis: an open and versatile identification key format. https://doi.org/10.1101/2022.05.26.493630 (May 2022). Preprint.

Declaration of contributions

Paper I

Wouter Koch contributed to conceptualization, data annotation, and text revisions. Caitlin P. Mandeville has led conceptualization, data annotation, analysis and visualization, writing. Erlend B. Nilsen and Anders G. Finstad have both contributed to conceptualization and text revisions.

Paper II

Wouter Koch led conception, wrote the code, executed the experiment and analyzed the results, wrote the main manuscript text and prepared all figures. Laurens Hogeweg contributed to the code. Both Laurens Hogeweg, Erlend B. Nilsen and Anders G. Finstad contributed with substantial feedback throughout the entire writing process and reviewed the final manuscript.

Paper III

Wouter Koch led conception, experimental design, code, analysis, and writing. Laurens Hogeweg contributed to the code, and all text revisions. Erlend B. Nilsen contributed to the conception, and text revisions. Robert B. O'Hara contributed to the analyses, and the text revisions. Anders G. Finstad contributed to the conception, the analyses, and text revisions.

Paper IV

Wouter Koch led conception, format design and code, writing, and visualization. Hallvard Elven has contributed to the conception, revision of the format design and code, and manuscript revisions. Anders G. Finstad has contributed to the manuscript revisions.

Introduction

Nature, and with it humanity, faces multiple crises at once. Climate change, the biodiversity crisis, mass extinction, habitat fragmentation and -loss, pollution, overexploitation, invasive alien species, etc., are all existential threats that need addressing[3–5]. While the effects on a larger scale are becoming painfully obvious, trends and the effects of counter measurements need to be monitored on a much more detailed scale, so conservation efforts can be guided and adjusted[6–8]. Such analyses are quite data hungry, and large amounts of data across space, time and taxonomy are needed to get a sufficiently fine grained view on status and trends in nature. For this, we need new ways of collecting data, whilst ensuring its quality, and to make sure that whatever data exists is available so it can be utilized to the fullest.

Citizen science

Amateur volunteers have long played a role in biodiversity data collection, and recently online repositories and collection platforms have begun to truly unlock the potential such "citizen scientists" can provide[9]. Something similar has happened before, when the invention of the steam powered printing press enabled a much broader dissemination of scientific literature in the 19$^{\text{th}}$ century[10]. This brought with it a rapid growth of commercial journals that were far more egalitarian and open to what was then called "low science", conducted by others than the traditional scholarly elite. While certainly not to everyone's taste, people like Charles Darwin were enthused by the broader societal contribution to scientific knowledge, with Darwin lamenting the absence of such contributions in the 'foreign periodicals' of the time; "a great loss it has always appeared to me"[10].

There are a great number of different definitions for what constitutes citizen science[11], which makes drawing direct historical parallels difficult. However, in the field of biology, one may argue that citizen science has had a prominent role from the start. Before biology or natural history were fields one could study in their own right, they were often the domain of naturalists who had the time and means to dedicate themselves to their interest in nature in their spare time. Many great early biologists, as we would now call them, stem from this tradition of "gentleman scientists" (those that were afforded this luxury were indeed almost exclusively men). Carl Linnaeus[12], regarded as the "father of taxonomy", was the son of a minister and amateur botanist. With young Carl being much more interested in going out in the field to look at plants than studying, a scholarly life was seen as an unlikely path, and he was sent on his way to be educated as a priest. Guided by several tutors that shared his interest in botany, he studied medicine, a field that included botany at the time and in which he ultimately became a professor. Gregor Mendel[13], the father of modern genetics, chose to become a monk at least in part because it would grant him the time and means to study and conduct

experiments. While educated and working as a priest and a teacher, he is mainly remembered for his groundbreaking work on heredity. Charles Darwin[14] came from a wealthy family and was sent off to study medicine. He neglected his studies which he found dull, instead spending time on taxidermy, beetle collecting and the likes. Lacking progress in his studies in medicine, his father sent him to study to become a clergyman instead. Charles again spent most of his time pursuing his interest in nature, but studied enough to graduate. Soon after his graduation he made his first voyage on the HMS Beagle, as an unpaid companion, and went on to become a renowned naturalist with an immensely influential legacy that lasts to this day.

Nowadays, having a hobby is no longer reserved for the wealthy and clerical. Many people of all walks of life find leisure in nature, and those with an interest in biology have the opportunity to privately study the taxa that interest them and become knowledgeable botanists, ornithologists, entomologists, etc.[9]. With this has come an extent of democratization of the study of biology. With a broad awareness of the challenges that face biodiversity, many citizen scientists wish to contribute to addressing conservational issues through their participation. The necessity of better insight into biodiversity trends, taking place over large geographical areas, comes with the need for large amounts of timely data from many places simultaneously. The expertise and engagement of a large community of amateur biologists provides an excellent fit for this, and citizen science as we understand it today was born.

The first time the term "citizen science" was used in its current meaning was in 1989[11,15]. By measuring the acidity of rain, citizens were helping the Audubon Society collect data that would be available as soon as possible, to better inform the political process. *"'Speed is also crucial to the Audubon Society's acid-rain campaign. Government studies sometimes withhold data for years', says Audubon vice president Robert San George, but 'the average citizen has trouble getting worked up about rain that fell a year and a half ago.' Audubon involves 225 society members from all 50 states in a 'citizen science' program that gets information out within five weeks. Volunteers collect rain samples, test their acidity levels, and report the results to Audubon headquarters, which releases a monthly national map of acid-rain levels. The information is used to lobby Congress."*[15]

It was thus the need for timely and openly accessible data that motivated the modern incarnation of the citizen science movement. Nearly 35 years on, numerous gaps in the existing data remain, and with it in our knowledge. Moreover, there is a need for continued monitoring to keep improving our understanding into the future. It is not feasible to have these data collected by scientists and environmental professionals alone; we simply cannot obtain the geographic and temporal scope one would ideally want with traditional methods. The engagement of citizen scientists helps address this. In the realm of observational data, citizen science now constitutes the vast majority of the data points, and this is only increasing. It is important that such data are available to all that need them for their research

or management task. Only in this way can existing data reach its potential and allow for as informed management decisions as possible.

Open availability of data

Even if we were somehow to acquire all the data that we need regarding the present, it takes a considerable temporal scope to detect changes in biodiversity statuses and trends over time. As it is very difficult to gather more data from the past, we need to do what we can to make sure such existing data are preserved, lest they be lost forever, and that anyone can access the data so they can be used wherever they are needed. Open access is also in line with the motivations of many citizen scientists, who collectively provide the largest share of the data. To ensure the maximum impact for management and research, access to observational data should not be restricted by one's budget, having the right affiliations or nationality, nor possession of advanced computing skills.

A common and useful framework to assess if data are truly open is to regard them within the framework of the FAIR concept; data need to be both Findable, Accessible, Interoperable, and Reusable[16]. If any of these prerequisites are not fulfilled, data reuse is hampered and its full societal potential is not achieved. There are numerous ways in which data can fail to be fully FAIR, with obstacles ranging from legal to practical in nature. There are licenses, platforms, data standards and guides to help individuals and institutions to ensure their data are ready and available for reuse, but there is still a need for added awareness regarding this issue. From curricula to research funding, open data must be not only required as an integral part of reproducible science, but be something for which time and resources are made available, and acknowledged as fullworthy, citable fruits of labor.

Once data is openly available, it is important to be aware of the fact that collecting observational data, and citizen science data collection in particular, brings with it a number of biases. Whenever observational data is collected by human observers, there will be behavioral, cultural, logistical, etc. reasons for taxa to either be observed and reported, or not. This has an effect on the nature and amounts of data that are available, and with it their value for science and management. When drawing conclusions from these data, it is vital to be aware of the ways in which the collection process is favoring certain times, places, or taxa over others, and account for these where possible. On the data collection end, awareness of biases can help guide the design of new protocols and systems that reduce them.

Taxonomic bias

Not all of the 1.8 million species currently described by science are reported as much as others. There are valid reasons for this to be the case; not every

species is as abundant, and if data were an unbiased representation of species' abundances and distributions, it is reasonable to expect that common species are more abundant in the data too. If this was the case, however, Antarctic krill (*Euphausia superba*) would be one of the most frequently reported species, but it is not[17]. There are many other factors that govern how likely individuals of a species are to be reported, all contributing to taxonomic biases.

In order to be reported, a taxon has to occur within the vicinity of an observer, who then has to spot it, recognize it, and report it. Each of these steps are non-random events whose likelihoods contribute to the taxonomic bias of the data. To use the example of the Antarctic krill, it occurs in deep, icy cold waters around Antarctica, a region and habitat that few humans venture into, so there are relatively few opportunities for humans and this species to cross paths. When this does happen, it has to be spotted by the would-be observer. Being transparent and only a few centimeters in size does not help in this regard. Then, once observed, the observer has to possess some knowledge of which species it is, or have access to the tools to do so. While most people would be able to recognize it as a shrimp-like crustacean, it requires more knowledge to recognize it as a species of krill. In order to identify it to the species level, one additionally has to be aware of the quite subtle differences with other species such as Antarctic coastal krill (*Euphausia crystallorophias*). Finally, the observer needs to be willing to put in the required effort to record and report the occurrence as an observation if it is to become part of the data corpus.

Thus, the geographic distribution, habitat, and biology of a species all impact how likely it is to be encountered and observed by a human. From that point, societal and individual preferences influence how likely a person is to know the species and take the time to report it. Knowledge on groups like titmice and other common garden birds is far greater than that on krill, as is the motivation to report sightings, leading to their relative over-representation in the available data. Such mechanisms are especially prevalent in citizen science, but not exclusive to it; scientific research and funding also favors taxa with a greater societal popularity.

There are ways to try to address these biases, which are being employed in citizen science and research more generally. Remotely operated vehicles, camera traps, acoustic monitoring, drones, canopy fogging, malaise traps, light traps, etc. can all help in encountering and detecting a more representative subset of species. Funding and outreach can help create research opportunities and raise awareness of the importance of reporting sightings. Tools like automated image recognition are essential in many of the efforts where large amounts of data are managed. They are also rapidly becoming more commonplace to aid in the identification process among citizen scientists, allowing for the reporting of a broader range of taxa than those that are typically known by the general public.

Image recognition

Pictures of species contain a lot of information, in the form of pixels with different red, green and blue values. For a person it can be trivial to recognize a depiction of something sufficiently familiar and distinguishable, but automating this feat has long proven to be an insurmountable challenge. One cannot define which pixel values correspond to which species, for example, as no two pictures are the same and any differences in subject placement, angle, lighting, etc. will result in a completely different image on the pixel level. The key lies in pattern recognition, a task for which the human brain is exceptionally well equipped.

In recent years, many technological improvements have been achieved that allow computers to better detect patterns in data in general, and pictures more specifically. In order to learn to distinguish meaningful patterns in the data from those patterns that do not carry any information for the task at hand, a lot of different examples in the form of training data are needed, and a lot of computing power to examine these training data in enough detail to find the most informative patterns. While new methods have contributed to the much improved capabilities of automated image recognition, they are mostly due to the availability of large amounts of data in the current information age, together with the rapid expansion of computing power, most notably in graphics processing units.

Applying image recognition can be done by using a convolutional neural network (CNN)[18]. In neural networks in general, the input forms a list or matrix of (input) values (e.g. the pixel values in the image). Based on these values, a new set of values is calculated. Each such calculation is the result of positive and negative relationships with multiple input values, analogous to how neurons in the brain inhibit or excite one another upon firing. Usually, many layers of newly calculated values are chained, where each layer informs the next, until the last layer, the output layer, produces a result. In a species recognition example, the output value can represent a list of species, where the aim is to maximize the value of the position corresponding to the correct species, and to minimize all others. A CNN adds an operation of convolution to this process. Here, each calculation is done on a block of values, so that pixels next to each other are considered in context of one another, maintaining the 2D structure of a picture. The network can then calculate to which degree a certain pattern, such as a diagonal line, is present within that region. By systematically sliding (convoluting) the area of focus over the entire picture, the result is a map of the entire picture of where that pattern is present to which degree. This resulting map can again be used as input for a convolutional layer, which then looks for patterns within the pattern from the previous step. In each step, this process can be done for multiple patterns in parallel, so each next layer is looking for more intricate combinations of patterns based on the results of the layers that came before it. Ultimately, this allows the CNN to recognize combinations of patterns that are complex enough to represent entire species.

It is not viable for anyone to program a CNN to look for the right combinations of patterns, and draw sensible conclusions from this. Instead, a process of machine learning is used, where patterns and links between patterns are randomized at the beginning of the learning process. By running a picture of which the true species is known through the network, it is possible to calculate how the links between each layer can be adjusted to result in a slightly better result for this example picture. Through doing this many times for many different examples, this mechanism attempts to find the best possible configuration of values within the CNN, giving the best answers. In essence, the model has learned to react only to the patterns within pictures that are most informative for deciding which species is depicted.

Code to create and train such models is widely available, and rapidly becoming more powerful and user friendly. This has led to widespread adoption, in this context especially within citizen science. Citizen science can provide the large amounts of data machine learning models need, it has users that vary in their species literacy and who can often benefit from these tools, and the large amounts of data citizen science generates renders manual validation by experts of what is reported nearly impossible. All this makes these two concepts great fits for one another, so it is little surprising that more and more citizen science platforms now provide image recognition tools in some form.

But pictures can only assist to a certain point. Even a world leading expert (or perfect artificial recognition model) will not be able to tell what is depicted in some cases. A picture will often lack the necessary information because the characteristics one has to look at to reliably distinguish between species are simply not visible in it[19]. It is therefore not plausible or desirable to try and classify every observation using image recognition models alone, and there will always be a need for the kind of deep taxonomic knowledge found in experts.

Taxonomic impediment

It takes years to become knowledgeable in the field of a sufficiently large or complex taxon, and world leading authorities on taxa have often spent large portions of their careers studying thousands of specimens in detail to gather the knowledge needed. In modern careers and curricula, there is generally not enough financial stability and time to make such investments in taxonomic expertise. As a result, this knowledge is disappearing from the scientific community, leading to what has been coined a “taxonomic impediment”[20].

The traditional way to convey the skills needed to distinguish species is through identification keys. While time consuming and demanding to make, and a necessarily incomplete representation of the true breadth of expert knowledge, it remains an invaluable tool for the aspiring taxonomist, allowing them to “stand on the shoulders of giants” in their learning trajectory. To preserve this vital knowledge, the availability of good, usable keys is paramount.

Research objectives

There is a great need for large amounts of observational data, and citizen science is proving crucial in obtaining it. Improving the data stemming from these efforts, used by science and management, requires awareness of (and ideally solutions for) improved open access to existing data, taxonomic bias and reporting bias in pictures due to recognizability, especially when considered in the context of automated image recognition, and finally the preservation and storage of expert knowledge to help citizen scientists, as it is steadily becoming more rare.

A lot of effort goes into methods for working with citizen science data and how to deal with its shortcomings in terms of coverage, quality and biases. Meanwhile, we should not be complacent about these issues, and aim for improved methodology that minimizes these issues at the collection stage, and in other ways aim to improve the biodiversity data we increasingly base both research and management on.

> The main objective of this thesis is to quantify the aspects in which opportunistic citizen science data is lacking or biased, and to investigate how this can be addressed at the collection end of the data life cycle.

This objective is subdivided into 4 sub-objectives:

- review the extent in which existing observational data are currently shared openly by researchers that gather or otherwise obtain it
- evaluate how citizen science picture data of different orders contribute to improving new recognition models, and how this relates to the well documented taxonomic biases in data availability
- investigate how picture quality and general recognizability on a species level relates to recognition model performance and data availability within citizen science
- propose a data format that captures identification key knowledge, so that citizen scientists (among others) can get easy access to the knowledge needed to identify taxa that cannot be easily identified by other means

The value of data depends on its ability to be used. Data that are not available cannot be used in analyses, and analyses building on data that is not also made available cannot be scrutinized. It is therefore vital to publish data openly for multiple reasons if it is to be part of a scientifically sound process. Many researchers have discovered these data as a source of information in their work. In paper I, we examine to which degree this increase in availability and use leads to an increased tendency to share data.

One application with a strong interrelatedness to citizen science is automated image recognition. Citizen scientists can benefit a lot from algorithms that can aid

in identifying taxa from images. At the same time, the training of such algorithms requires amounts of data that can only be provided by citizen science. In paper II, we assess the unequal representation of taxa in Norwegian citizen science image data, and whether we need more of what we have the least of, or if there is a difference between the taxa, and thus the impact that a contribution by a citizen scientist can have in this respect.

Not every species is easily recognizable from a photograph, if at all. There may also be reasons for photographs to be of a different quality depending on the species, depending on the behavior of the species and the interest it receives from different parts of the citizen science community. These factors impact both the likelihood of an observation being accompanied by a picture, as well as the expected maximum recognition model performance. In paper III, we investigate these mechanisms in the data corpus, as well the consequences for how we expect future models to perform and the kind of data collection we aim for within citizen science.

Given the limitations in machine learning, and the impossibility to reliably identify many species from pictures in general, there will always be a need for taxonomic knowledge. Such knowledge is becoming more scarce as we are facing a taxonomic impediment, however. It is important to store such knowledge, as well as share it in a user-friendly way. Digital keys are a logical way to do this, and come with the added benefit of being able to represent knowledge that cannot be represented in paper form in a practical way, greatly enhancing the educational potential of digital keys. There is however no open format that takes advantage of the full breadth of possibilities or modern standards. We address this in paper IV, where we propose a data format for storing taxonomic knowledge in an open, modern, and flexible manner.

Summary of papers

Paper 1

Presence-only biodiversity data are increasingly relied on in biodiversity, ecology, and conservation research, driven by growing digital infrastructures that support open data sharing and reuse. Recent reviews of open biodiversity data have clearly documented the value of data sharing and there is growing recognition that this open sharing of biodiversity data is critical for advancing biodiversity research[21]. Some of the primary benefits of open biodiversity data include enhanced reproducibility of research[22]; making data available for reuse in new research applications[23]; enabling researchers to receive credit, in the form of citations, for their efforts producing and sharing data sets[24,25]; and minimizing the duplication of research effort, enabling researchers to prioritize new data collection that fills research gaps[26].

Many aspects of the sharing and reuse of openly accessible biodiversity data in the peer-reviewed literature have been characterized, including common research applications of open data, taxonomic and spatial trends in open data, persistence of data stored in open databases, and current citation practices for open data[26–30]. These studies make it clear that openly shared presence-only biodiversity data are foundational to a large body of biodiversity research. Still, many data go unshared. Earlier in the open data movement, it was widely recognized that open data formed just a small portion of the total biodiversity data known to exist[31–33]. But the current volume of presence-only data that are not openly shared, despite being presented and analyzed in the literature, is unknown.

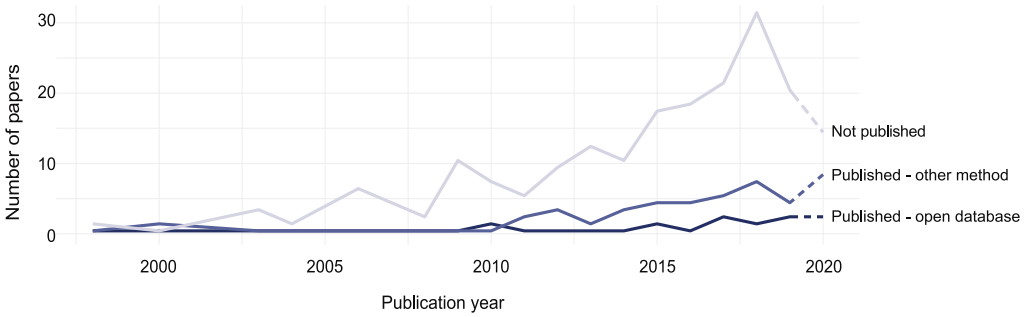


Figure 1: Over time, more and more data are shared openly, but much remains inaccessible.

We address this question by examining a broad cross section of the traditional peer-reviewed literature to assess the degree to which it promotes and implements open presence-only biodiversity data. Our goal is to provide insight into the current adoption of open data practices among users of presence-only biodiversity data in journals drawn from a variety of sources beyond open databases, in the indexed literature. We focus on how frequently researchers access open data relative to

data from other sources, how often they share newly generated or collated data, and trends in metadata documentation and data citation.

We find that the sharing of presence-only biodiversity data is overall increasing but that there is ample room for improvement in adherence to many data sharing best practices (figure 1). Biodiversity research commonly relies on presence-only data that are not openly available and neglects to make such data available, still limiting its value, reusability, and reproducibility.

Paper 2

Technological advances and data availability have enabled artificial intelligence-driven tools that can increasingly successfully assist in identifying species from images. Especially within citizen science, an emerging source of information filling the knowledge gaps needed to solve the biodiversity crisis, such tools can allow participants to recognize and report more poorly known species[34–36]. This can be an important tool in addressing the substantial taxonomic bias in biodiversity data, where broadly recognized, charismatic species are highly over-represented[37]. Meanwhile, the recognition models are trained using the same biased data, so it is important to consider what additional images are needed to improve recognition models.

We use the Species Observation Service[38], a large Norwegian citizen science project, as an example to investigate the nature of the bias in citizen science image data, and how this relates to the value of data for image recognition models. One way to evaluate this is by using the concept of Value of Information (VoI); "*the increase in expected value that arises from making the best choice with the benefit of a piece of information compared to the best choice without the benefit of that same information*"[39].

Considering training data for image recognition models in the VoI framework allows us to identify the most effective prioritization for improving recognition models. This method allows for a more sophisticated approach to data collection than simply adding more data for all taxa, or prioritizing taxa that are currently the most under-represented. First, we evaluate whether the biases found in observation data in general, regardless of source, are the same within citizen science observations with images, or if there are different biases that need to be taken into account. Then we train multiple image recognition models for different taxa, with a gradually increasing number of images per species, allowing us to quantify and compare the effects of adding more training data between taxa. Using these changes in performance, we estimate the VoI of adding training data for each taxon, relative to the amount of images that are currently available. Finally, comparing this VoI to the amount of over- or under-representation of these taxa, we demonstrate that mobilizing images with a higher VoI provides an alternative, data-driven approach to simply prioritizing images of the currently most under-represented taxa.

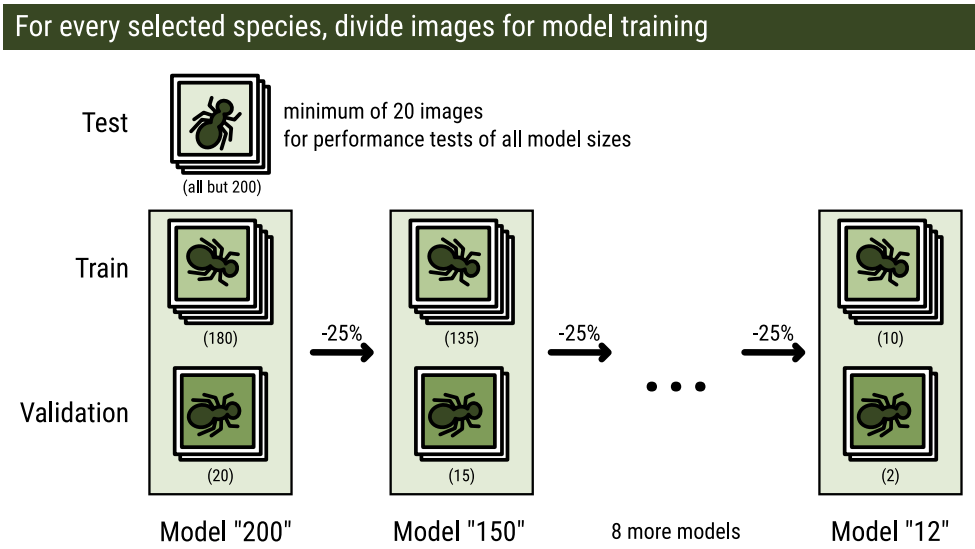


Figure 2: Data selection and subdivision. Each run is generated by selecting 17 species per order, using 200 images per species. For each subsequent model in a run, training and validation data are reduced by 25% (or slightly less than 25% if not divisible by 4).

We selected 12 orders in such a way that each order contained at least selected 17 species, where each such species had a minimum of 220 citizen science observations with at least one image. For each order, image recognition models were then trained using 200, 150, 113, 85, 64, 48, 36, 27, 21, 16 and 12 observations for training and validation, repeating this procedure 5 times with different subsets of 17 species, training a total of 660 models (figure 2). This enabled us to see how recognition performance increases per taxon as more images per species are available. Taking the curve of the performance increase then let us calculate the VoI of the taxon at any given point; the expected increase in recognition performance when adding one extra observation with images for each species in that taxon. Calculating the VoI for the current number of images available per species in the taxon (on average) provides a way to compare orders by how much a newly added observation with images to the currently available data would be worth in terms of recognition model improvement.

We found that, as is known for other data sets, there is a substantial taxonomic bias in the Norwegian citizen science data with images. The orders of "large colorful flying animals" like butterflies, birds and dragonflies are the orders that are most over represented in the data, with less conspicuous orders like flies, lichens and beetles trailing the list. When we however consider the current expected VoI per added observation with images for each order, we find that it is not strictly a matter of the most underrepresented orders benefiting the most from an addition to the data currently available. While there is a trend that such orders benefit more on average, by far the largest increase is found within the plant orders (figure

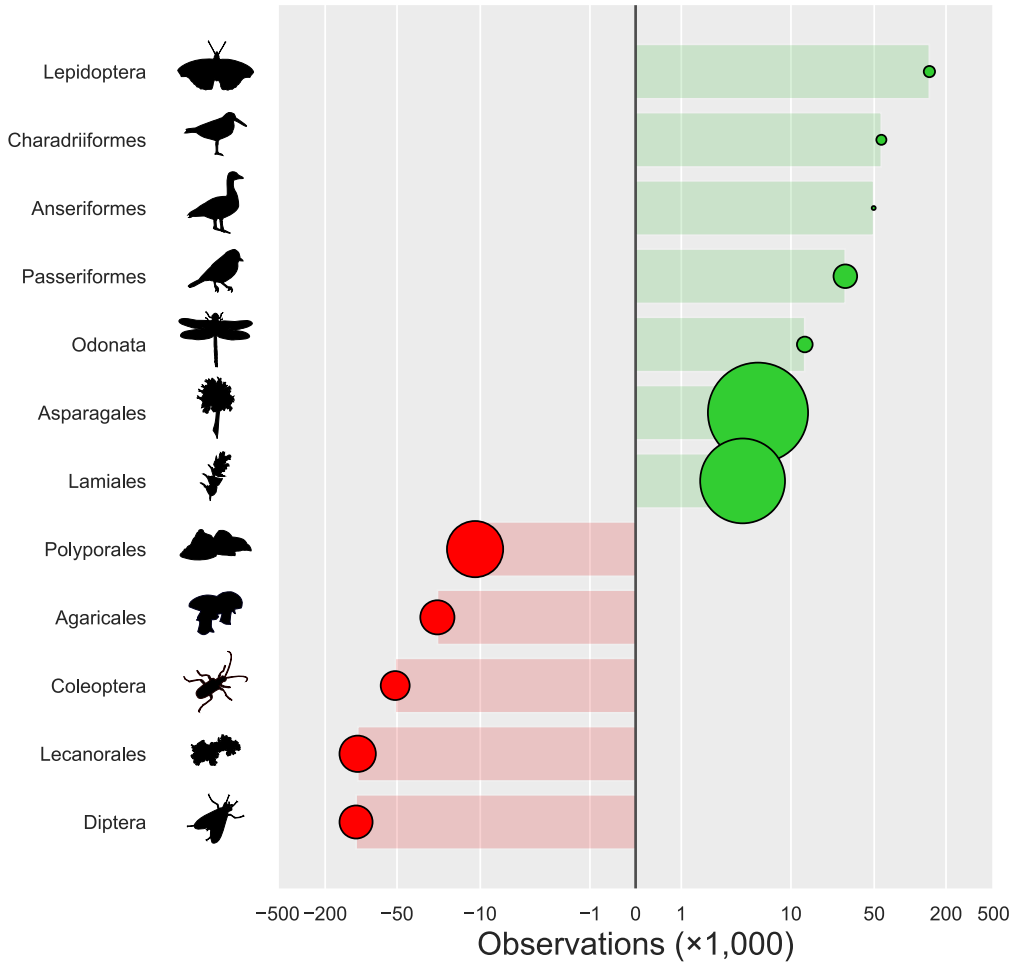


Figure 3: The over- (green) and under-representation (red) of species in citizen science image data. The areas of the circles represent the expected benefit for a recognition model when adding one more image to what is currently available.

3).

We thus demonstrate that a more informed decision is possible when choosing to focus on certain taxa for data collection aimed at improved recognition models. Prioritization of taxa for which to mobilize additional data can be informed by considering its expected VoI, rather than simply prioritizing those that are currently the most under-represented numerically. Note that this is no plea for deprioritizing data collection for such taxa in the context of citizen science as a whole. There are many areas of management and research that can benefit from additional data on taxa we predict will benefit less from additional images for recognition models, and ample reasons to mobilize data for other applications than image recognition.

Paper 3

Occurrence data are typically subject to spatial, temporal and taxonomic bias[37,40], and traditional manual methods of data collection are insufficient to gather the data volume needed, or address these biases.

Alternative methods[9,41,42] are being deployed to gather large amounts of data. With the increased output from initiatives like citizen science and camera-traps automating insect monitoring, manual management and quality control become infeasible. Automated image recognition tools for species identification are increasingly used to alleviate this[34–36,43]. Training image recognition models, however, also requires large amounts of pictures[44]. This creates a mutual reliance between large scale image data collection and image recognition models[45].

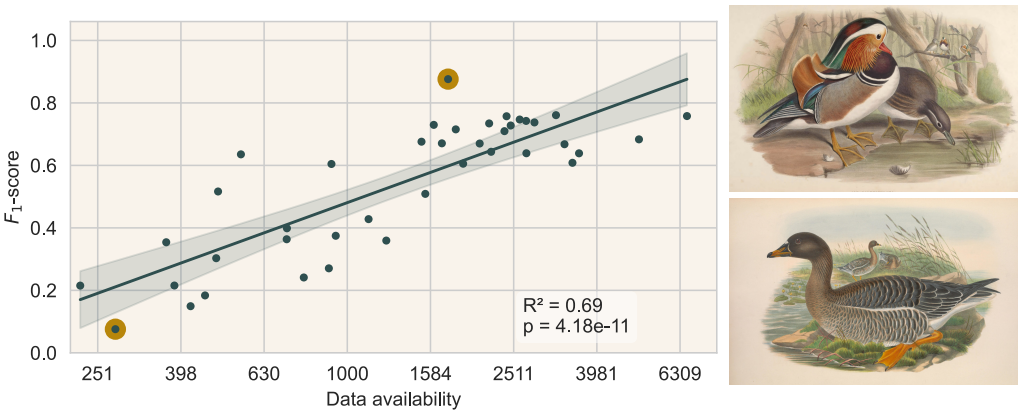


Figure 4: Example taxon where data availability is correlated with model performance for a species (even though the same amount of data is used for all species). The best and poorest performing species (highlighted and depicted) are examples where the former is much more distinct and recogntizable than the latter.

Visual identification of species is a complex task; while some species are unmistakable, many others are very challenging or even outright impossible to identify, regardless of picture quality[19]. Models are trained using training data reported and identified by humans, so species with low recognizability among humans may be underreported and be underrepresented in the training data. This means that recognition models are then being trained with data consisting mostly of pictures of species that are easier to recognize. In that case, training models on difficult species will be even harder, given their absence from the training data, which comes in addition to their already more challenging nature in terms of recognition.

To evaluate the existence of this possible reporting bias and its consequences, we evaluated how data availability, picture quality, biological traits and citizen science data collection differs across species, and how these differences relate to recognition model performance. We find evidence for a “recognizability bias”,

where species that are more readily identified by humans and recognition models alike are more prevalent in the available image data (figure 4). This pattern is present across multiple taxa, and does not appear to relate to a difference in picture quality, biological traits, or data collection metrics other than recognizability.

Paper 4

Research and nature management are facing a "taxonomic impediment", where taxonomic knowledge is gradually disappearing from the scientific community[20]. At the same time, it is clear that these skills are strongly needed in biodiversity monitoring for management and conservation, especially when carried out by citizen scientists.

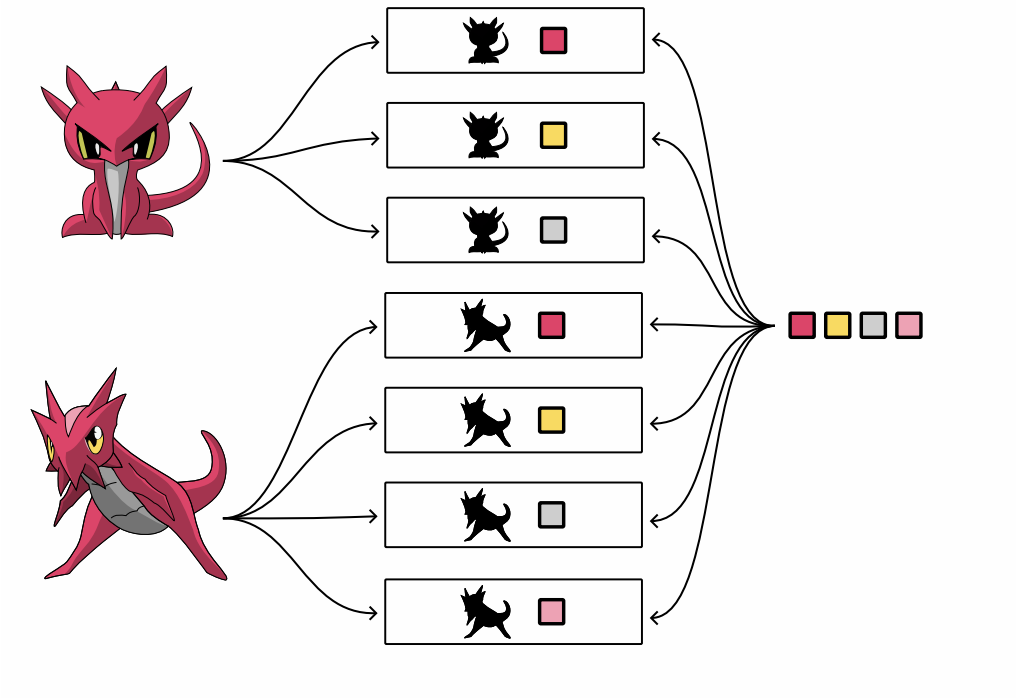


Figure 5: The core structure of Clavis: taxa and their characteristics are connected to one another through a collection of "statements".

Formalizing the required knowledge in the form of digital identification keys is one way of making such knowledge more available for professional and amateur observers of biodiversity. In paper 4 we describe Clavis, a modern open format for capturing knowledge required for taxon identification through digital keys, allowing for a level of detail beyond that of any current key format[46,47]. The core of a Clavis-compliant identification key are its statements which describe

which values a taxon has for which characteristic (figure 5). The format allows for many different types of metadata to support complicated logic, exceptions, documentation, links, geography, multimedia, provenance, etc. We exemplify each concept using Pokémon as a fictional taxonomic group[48], to ensure that no taxonomic disputes distract from the exemplified concepts while adhering to a predefined taxonomy with sufficient complexity.

Discussion

Open access to high quality biodiversity occurrence data is key to many themes in biodiversity research and conservation[49–51], and efforts to increase the open sharing of biodiversity data will be critical. Recent trends toward increased sharing of presence-only biodiversity data are a cause for optimism. There is a recent increase in the proportion of articles that produce open data, authors make the effort to credit original data providers as best they can, and citizen science data is shared at above-average rates. Still, there is a great deal of work to be done in normalizing the use of best practices in data access, documentation, citation, and sharing. Researchers generally feel positively toward reusing and sharing data, but are uncertain about best practices and concerned about credit and incentives[52–54]. Addressing these issues will be essential to meet challenges associated with the growing biodiversity crisis and to support a growing need for biodiversity assessment, monitoring, and conservation.

In gathering new data, image recognition tools play an important role in maintaining the quality of the large amounts of biodiversity data science and management require. Training these models requires substantial amounts of data, and as more images are collected through citizen science, recognition models can be steadily improved. Meanwhile, we find evidence that both the data that are currently available, and the potential informational value of added data are not taxonomically neutral.

With the more widespread use of image recognition models as both a user tool and a mechanism for quality control, it is time to view images as data in and of themselves, rather than only documentation of occurrences. Such a shift calls not only for conscious choices when it comes to the value of information in images, but increased implementation of data practices such as persistent storage, metadata standardization and the other FAIR data principles[55] to enable more apt usage of image data for current and novel applications.

Still, there are inherent limits to what can be identified from a picture, and identification tools are needed that rely on more than just pixel information. Models that take into account season, location, sound, etc. can be especially beneficial for difficult species. But ultimately, there is no substitute for the taxonomic knowledge of experts. Preserving this knowledge, and making it available in the form of identification keys, is vital. The open exchange of taxonomic knowledge, unambiguously captured with as much of the auxiliary details needed for its application, is essential for the preservation of invaluable, increasingly elusive knowledge. Such tools can help greatly to more reliably identify challenging species, in tandem with automatic identification. The data quality benefits ultimately feed back into the areas where such data are used, from research and spatial distribution models to the decision making processes related to the biodiversity crisis in a changing world.

References

1. Schwartz, M. A. The importance of stupidity in scientific research. *Journal of Cell Science* **121,** 1771–1771. ISSN: 0021-9533. https://doi.org/10.1242/jcs.033340 (June 2008).
2. Muff, S., Nilsen, E. B., O'Hara, R. B. & Nater, C. R. Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution* **37,** 203–210. https://doi.org/10.1016/j.tree.2021.10.009 (Mar. 2022).
3. IPBES. *Thematic assessment of the sustainable use of wild species of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (eds Fromentin, J.-M. *et al.*) https://doi.org/10.5281/zenodo.6448567 (2022) (IPBES Secretariat, Bonn, Germany, July 2022).
4. IPCC. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Pörtner, H.-O. *et al.*) (Cambridge University Press. In Press, 2022).
5. Secretariat of the Convention on Biological Diversity. *Global Biodiversity Outlook 5* Sept. 2020.
6. Xu, H. *et al.* Ensuring effective implementation of the post-2020 global biodiversity targets. *Nature Ecology & Evolution* **5,** 411–418. https://doi.org/10.1038/s41559-020-01375-y (Jan. 2021).
7. Wetzel, F. T. *et al.* Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in Europe. *Biological Conservation* **221,** 78–85. https://doi.org/10.1016/j.biocon.2017.12.024 (May 2018).
8. Scholes, R. J. *et al.* Toward a Global Biodiversity Observing System. *Science* **321,** 1044–1045. https://doi.org/10.1126/science.1162055 (Aug. 2008).
9. Silvertown, J. A new dawn for citizen science. *Trends in Ecology & Evolution* **24,** 467–471. https://doi.org/10.1016/j.tree.2009.03.017 (Sept. 2009).
10. Dawson, G., Lintott, C. & Shuttleworth, S. Constructing Scientific Communities: Citizen Science in the Nineteenth and Twenty-First Centuries. *Journal of Victorian Culture* **20,** 246–254. https://doi.org/10.1080/13555502.2015.1022053 (Mar. 2015).
11. Haklay, M. *et al.* in *The Science of Citizen Science* 13–33 (Springer International Publishing, 2021). https://doi.org/10.1007/978-3-030-58278-4_2.
12. Bolton, S. K. *Famous Men of Science* https://gutenberg.org/ebooks/35489 (Thomas Y. Crowell & Co., New York, 1889).

13. Weiling, F. Historical study: Johann Gregor Mendel 1822–1884. *American Journal of Medical Genetics* **40,** 1–25. https://doi.org/10.1002/ajmg.1320400103 (1991).

14. Darwin, C. *The Autobiography of Charles Darwin* https://gutenberg.org/ebooks/2010 (John Murray, London, 1887).

15. Kerson, R. *Lab for the environment* 1989.

16. Wilkinson, M. D. *et al. The FAIR Guiding Principles for scientific data management and stewardship* Mar. 2016. https://doi.org/10.1038/sdata.2016.18.

17. GBIF.org. *GBIF Occurrence Download* 2022. https://doi.org/10.15468/dl.zv45s2.

18. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org. http://www.deeplearningbook.org (MIT Press, 2016).

19. Lukhtanov, V. A. Species Delimitation and Analysis of Cryptic Species Diversity in the XXI Century. *Entomological Review* **99,** 463–472. https://doi.org/10.1134/s0013873819040055 (July 2019).

20. Engel, M. S. *et al.* The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society* **193,** 381–387. https://doi.org/10.1093/zoolinnean/zlab072 (Sept. 2021).

21. Farley, S. S., Dawson, A., Goring, S. J. & Williams, J. W. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience* **68,** 563–576. https://doi.org/10.1093/biosci/biy068 (July 2018).

22. Alston, J. M. & Rick, J. A. A beginner's guide to Conducting reproducible research. *Bulletin of the Ecological Society of America* **102,** 1–14. https://doi.org/10.1002/bes2.1801 (2021).

23. Chawinga, W. D. & Zinn, S. Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research* **41,** 109–122. https://doi.org/10.1016/j.lisr.2019.04.004 (2019).

24. Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q. & Bourne, P. E. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution* **28,** 454–461. https://doi.org/10.1016/j.tree.2013.05.002 (2013).

25. Brown, R. F. The importance of data citation. *BioScience* **71,** 211–211. https://doi.org/10.1093/biosci/biab012 (2021).

26. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7.** https://doi.org/10.1038/s41598-017-09084-6 (Aug. 2017).

27. Escribano, N., Galicia, D. & Ariño, A. H. The tragedy of the biodiversity data commons: a data impediment creeping nigher? *Database* **2018.** https://doi.org/10.1093/database/bay033 (2018).

28. Ball-Damerow, J. E. *et al.* Research applications of primary biodiversity databases in the digital age. *PloS one* **14,** e0215794. https://doi.org/10.1371/journal.pone.0215794 (2019).

29. Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B. & Schigel, D. Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences* **118.** https://doi.org/10.1073/pnas.2018093118 (2021).

30. Luo, M. *et al.* The use of Global Biodiversity Information Facility (GBIF)-mediated data in publications written in Chinese. *Global Ecology and Conservation* **25,** e01406. https://doi.org/10.1016/j.gecco.2020.e01406 (2021).

31. Ariño, A. H. Approaches to estimating the universe of natural history collections data. *Biodiversity informatics* **7.** https://doi.org/10.17161/bi.v7i2.3991 (2010).

32. Amano, T., Lamming, J. D. & Sutherland, W. J. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience* **66,** 393–400. https://doi.org/10.1093/biosci/biw022 (2016).

33. Peterson, A. T., Asase, A., Canhos, D. A. L., de Souza, S. & Wieczorek, J. Data leakage and loss in biodiversity informatics. *Biodiversity Data Journal.* https://doi.org/10.3897/BDJ.6.e26826 (2018).

34. Weinstein, B. G. A computer vision for animal ecology. *Journal of Animal Ecology* **87** (ed Prugh, L.) 533–545. https://doi.org/10.1111/1365-2656.12780 (Nov. 2017).

35. Wäldchen, J., Rzanny, M., Seeland, M. & Mäder, P. Automated plant species identification—Trends and future directions. *PLOS Computational Biology* **14** (ed Bucksch, A.) e1005993. https://doi.org/10.1371/journal.pcbi.1005993 (Apr. 2018).

36. Ceccaroni, L. *et al.* Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. *Citizen Science: Theory and Practice* **4.** https://doi.org/10.5334/cstp.241 (2019).

37. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7.** https://doi.org/10.1038/s41598-017-09084-6 (Aug. 2017).

38. The Norwegian Biodiversity Information Centre. *Norwegian Species Observation Service* 2022. https://doi.org/10.15468/zjbzel.

39. Keisler, J. M., Collier, Z. A., Chu, E., Sinatra, N. & Linkov, I. Value of information analysis: the state of application. *Environment Systems and Decisions* **34,** 3–23. https://doi.org/10.1007/s10669-013-9439-4 (Apr. 2013).

40. Boakes, E. H. *et al.* Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *PLoS Biology* **8,** e1000385. https://doi.org/10.1371/journal.pbio.1000385 (June 2010).

41. Hansen, O. L. P. *et al.* Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and Evolution* **10,** 737–747. https://doi.org/10.1002/ece3.5921 (Dec. 2019).

42. Kirkeby, C. *et al.* Advances in automatic identification of flying insects using optical sensors and machine learning. *Scientific Reports* **11.** https://doi.org/10.1038/s41598-021-81005-0 (Jan. 2021).

43. Christin, S., Hervet, É. & Lecomte, N. Applications for deep learning in ecology. *Methods in Ecology and Evolution* **10** (ed Ye, H.) 1632–1644. https://doi.org/10.1111/2041-210x.13256 (July 2019).

44. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org (MIT Press, 2016).

45. Lotfian, M., Ingensand, J. & Brovelli, M. A. The Partnership of Citizen Science and Machine Learning: Benefits, Risks, and Future Challenges for Engagement, Data Collection, and Data Quality. *Sustainability* **13,** 8087. https://doi.org/10.3390/su13148087 (July 2021).

46. Dallwitz, M. J. A general system for coding taxonomic descriptions. *TAXON* **29,** 41–46. https://doi.org/10.2307/1219595 (Feb. 1980).

47. Identic Pty Ltd. *Lucidcentral.org* https://www.lucidcentral.org (2022).

48. Fandom Community. *Pokémon GO Wiki* https://pokemongo.fandom.com/wiki/List_of_Pok%C3%A9mon (2022).

49. Hochkirch, A. *et al.* A strategy for the next decade to address data deficiency in neglected biodiversity. *Conservation Biology* **35,** 502–509. https://doi.org/10.1111/cobi.13589 (2021).

50. Nakagawa, S. *et al.* A new ecosystem for evidence synthesis. *Nature Ecology & Evolution* **4,** 498–501. https://doi.org/10.1038/s41559-020-1153-2 (2020).

51. Callaghan, C. T. *et al.* Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience* **71,** 55–63. ISSN: 0006-3568. https://doi.org/10.1093/biosci/biaa131 (Nov. 2020).

52. Ross-Hellauer, T., Deppe, A. & Schmidt, B. Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. *PloS one* **12,** e0189311. https://doi.org/10.1371/journal.pone.0189311 (2017).

53. Tenopir, C. *et al.* Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PloS one* **15,** e0229003. https://doi.org/10.1371/journal.pone.0229003 (2020).

54. Soeharjono, S. & Roche, D. G. Reported individual costs and benefits of sharing open data among Canadian Academic Faculty in ecology and evolution. *BioScience* **71,** 750–756. https://doi.org/10.1093/biosci/biab024 (2021).

55. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3.** https://doi.org/10.1038/sdata.2016.18 (Mar. 2016).

Paper I

BIOSCIENCE, VOLUME 71, ISSUE 11, NOVEMBER 2021, PAGES 1128-1147, DOI:10.1093/BIOSCI/BIAB072

Open Data Practices among Users of Primary Biodiversity Data

Caitlin P. Mandeville[1] Wouter Koch[1,2] Erlend B. Nilsen[3] Anders G. Finstad[1]

1. Department of Natural History, Norwegian University of Science and Technology, Trondheim, Norway; 2. Norwegian Biodiversity Information Centre, Trondheim, Norway; 3. Faculty of Biosciences and Aquaculture, Nord University, Steinkjer, Norway

Published: 18 August 2021

Abstract

Presence-only biodiversity data are increasingly relied on in biodiversity, ecology, and conservation research, driven by growing digital infrastructures that support open data sharing and reuse. Recent reviews of open biodiversity data have clearly documented the value of data sharing, but the extent to which the biodiversity research community has adopted open data practices remains unclear. We address this question by reviewing applications of presence-only primary biodiversity data, drawn from a variety of sources beyond open databases, in the indexed literature. We characterize how frequently researchers access open data relative to data from other sources, how often they share newly generated or collated data, and trends in metadata documentation and data citation. Our results indicate that biodiversity research commonly relies on presence-only data that are not openly available and neglects to make such data available. Improved data sharing and documentation will increase the value, reusability, and reproducibility of biodiversity research.

Keywords: applied ecology, biodiversity, informatics, monitoring and mapping, publication practices

Biodiversity data are increasingly made openly available, facilitated by extensive digital infrastructures that support data standardization and publication[1–3]. There is growing recognition that this open sharing of biodiversity data is critical for

advancing biodiversity research[1]. Some of the primary benefits of open biodiversity data include enhanced reproducibility of research[4]; making data available for reuse in new research applications[5]; enabling researchers to receive credit, in the form of citations, for their efforts producing and sharing data sets[6,7]; and minimizing the duplication of research effort, enabling researchers to prioritize new data collection that fills research gaps[8]. As data sharing continues to become normalized, best practices have developed for the sharing of biodiversity data[9]. The FAIR data principles, for instance, outline four key attributes of effectively shared data: findable, accessible, interoperable, and reusable[10]. Specific practices have been developed to implement biodiversity data sharing in accordance with FAIR data principles. For example, global data aggregators such as the Global Biodiversity Information Facility (GBIF) provide a central location for aggregated data sets, ensuring that they will be findable and accessible[11], and standardization schemes such as Darwin Core provide a mechanism for researchers to improve interoperability[12]. Such innovations support extensive data reuse; for example, the GBIF currently enables integrated data searches of nearly 1.7 billion species records from diverse sources around the world and has facilitated data reuse in thousands of publications[13].

Although any type of data can be openly shared, the biodiversity data type most readily associated with open data sharing is presence-only occurrence data[2,14–16]. Presence-only data consist of the taxonomic identification and location of an organism, often with the time of observation but without further information about species abundance, sampling design, or sites at which the species was not observed. The quantity of presence-only data aggregated in open biodiversity data repositories is immense and continuing to grow rapidly[17,18]. This growth has been driven in large part by two simultaneous trends: the increasing popularity of citizen science platforms through which the public submit opportunistic observations to centralized databases[19–21] and the digitization and aggregation of historical records and museum specimens[22–25]. The growing volume of openly shared presence-only data is also driven by characteristics of the data type itself: It is relatively simple and is easily standardized within existing best practices for data sharing[2]. Presence-only occurrence data now offer greater spatial, temporal, and taxonomic coverage on a global scale than other biodiversity data types and are often less costly and time intensive to collect[26,27].

As presence-only biodiversity data have grown in volume and accessibility, they have become increasingly common in biodiversity research[13,17]. The open availability of massive modern and historical biodiversity data sets has contributed to a wide range of research areas, including ecology, biogeography, global change, and conservation[13,18,28]. But the analysis of presence-only data is not without challenges; both historical and modern presence-only data are associated with limitations and biases that are distinct from other data types, both because of the lack of absence data and also because of the opportunistic collection process frequently associated with presence-only data[28–34]. Further biases, errors, and

limitations can be introduced in the processes of data preparation, publishing, and long-term maintenance[35,36], including the issues of data leakage[17] and data obsolescence[37]. In response to these challenges, the growing application of presence-only data has been paralleled by an explosion of innovation in approaches to assess and improve both data accessibility and quality[18] and also analysis methods that account for the specific limitations associated with this data type[38,39]. As the development of analysis approaches for presence-only data continues, there is broad consensus that the documentation of metadata that details the study protocol, including information about sampling design or effort, allows for greater inference and also greater data reuse and reproducibility of analyses[39–42]. Open biodiversity data repositories commonly encourage the publishing of metadata[43], but in practice the quality and amount of documented metadata varies widely[2,17,44].

Although presence-only biodiversity data are reported and analyzed extensively in the traditional peer-reviewed literature, they are not restricted to it. In particular, authors who publish or access openly accessible biodiversity data may be more likely to seek out alternative outlets for research publication, such as preprint servers and journals with novel publishing models, because of their emphasis on free sharing of scientific information. Furthermore, biodiversity data are likely reported and analyzed often in gray literature and conference proceedings. Still, because a great deal of biodiversity data are reported and analyzed in the traditional peer-reviewed literature, it is important to understand the role that this literature plays in either facilitating or hindering the open sharing of biodiversity data. In this review we consider the extent of and barriers to the adoption of open data sharing practices within the traditional peer-reviewed literature, represented by the set of journals indexed by the Web of Science Core Collection.

Many aspects of the sharing and reuse of openly accessible biodiversity data in the peer-reviewed literature have been characterized, including common research applications of open data, taxonomic and spatial trends in open data, persistence of data stored in open databases, and current citation practices for open data[8,13,18,45,46]. These studies make it clear that openly shared presence-only biodiversity data are foundational to a large body of biodiversity research. Still, many data go unshared. Earlier in the open data movement, it was widely recognized that open data formed just a small portion of the total biodiversity data known to exist[17,20,47]. But the current volume of presence-only data that are not openly shared, despite being presented and analyzed in the literature, is unknown. The concept of data sources and sinks can be helpful to conceptualize this issue; publication approaches that generate or perpetuate openly shared data can act as sources for continued data reuse, whereas publication approaches that entail a single use of data with no means for open access or reuse can be thought of as data sinks.

In the present article, we examine a broad cross section of the traditional peer-reviewed literature to assess the degree to which it serves as a source or sink for open presence-only biodiversity data. Our goal is to provide insight

into the current adoption of open data practices among users of presence-only biodiversity data in journals indexed by the Web of Science Core Collection. To our knowledge, this is the first review of open data practices to be broadly defined by the presence-only data type, rather than by a particular type of data source, such as open databases. We focus on the following questions: How commonly does research published in articles indexed by the Web of Science Core Collection rely on presence-only data from open sources, and how commonly does it rely on data that are newly generated or compiled from other sources? To what extent do articles indexed by the Web of Science Core Collection serve as a data source for open presence-only biodiversity data; that is, are newly generated or compiled data made openly available, and are open data analyzed, documented, and cited in a way that supports continued reuse?

We identify both successes and challenges in the open sharing of presence-only biodiversity data, finding that the sharing of presence-only biodiversity data is overall increasing but that there is ample room for improvement in adherence to many data sharing best practices. We compare these findings with those of other recent reviews of the biodiversity literature, discussing trends that may be distinct to the presence-only data type, as well as new patterns that may be emerging within open data sharing practices. Because presence-only data are the biodiversity data type most commonly associated with open data sharing, they can serve as an early indicator to illustrate the developing state of data sharing more broadly in the related fields of biodiversity, ecology, and conservation. Therefore, our characterization of current practices in presence-only data sharing can illuminate successes, challenges, and barriers to the adoption of data sharing practices that may be of growing relevance to the greater biodiversity research community.

Review of the presence-only biodiversity data literature

We searched the Web of Science Core Collection to target all scholarly articles that report on the application of presence-only biodiversity occurrence data. Our search targeted articles whose titles, abstracts, or keywords contained any of 31 terms commonly used in the literature to indicate presence-only data as well as any of 5 terms used to indicate biodiversity (box 6). We screened the abstracts of all returned articles and retained those that demonstrated the analysis or reporting of presence-only occurrence data. After screening, a total of 2151 articles were included in the review (see the extended methods description in supplemental file S1). Data management and bibliometric summary statistics were conducted in part with the bibliometrix package in R[48].

To identify broad trends in applications of presence-only data, we classified all included articles into three topic clusters using latent dirichlet allocation (LDA) topic modeling. LDA topic modeling uses word associations within a corpus to identify topic clusters and assigns documents to the topic clusters on the basis of

```
(((TS = ("presence-only" OR "presence only" OR "opportunistic
    observation*" OR "opportunistic species observation*" OR
    "opportunistic occurrence*" OR "opportunistic distribution*" OR
    "opportunistic species occurrence*" OR "opportunistic species
    distribution*" OR "pseudo-absence*" OR "pseudoabsence*" OR
    "inferred absence*" OR "presence-background" OR "presence
    background" OR "citizen science" OR "community science" OR
    "participatory science" or "ad hoc data" OR "ad hoc collection"
    OR "ad hoc method*" OR "incidental data" OR "incidental
    sighting*" OR "incidentally collected" OR "geographic one-class
    data" OR "incidental detection*" OR "opportunistic detection*" OR
    "primary biodiversity data*" OR "occurrence record*" OR "atlas
    data" OR "unstructured occurrence data" OR "unstructured species
    observation" OR "unstructured biodiversity data")) AND (TS =
    ("distribution" OR "species" OR "biodiversity" OR "habitat*" OR
    "niche*")))
AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)
Indexes = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan
    = All years
```

Box 1: The search string used to query the Web of Science Core Collection to obtain literature.

word frequency within each document[49]. We classified each document on the basis of the words in the abstract and title. LDA topic modeling requires the desired number of clusters to be defined, so to select a number of topic clusters we conducted LDA analysis six times, each time producing a different number of clusters ranging from three to eight. We used two criteria to select the number of clusters in our final topic model: First, we assessed the clusters for lack of redundancy in an ordination of all articles by their highest rated topic classification, and, second, we assessed the redundancy and interpretability of the sets of most highly weighted words in each set of clusters[49,50] (see supplemental file S2). The modeling iteration that produced three topic clusters was least redundant and most interpretable. The topic clusters were assigned descriptive names on the basis of the words most characteristic of each cluster: methodological articles were characterized by terms related to the application and assessment of analysis methods; applied articles were characterized by terms related to topics in biodiversity science, conservation, and related fields; and records articles were characterized by terms related to the collection and reporting of occurrence data (figure 6). Topic modeling was conducted with the revtools package in R[49].

A subset of 300 articles randomly selected from the included articles was read in full and coded according to a standardized data sheet (see supplemental files S3 and S4). The 300-article subset was representative of the full data set in terms of publication year and topic area (figure 7). For each article read in full, we

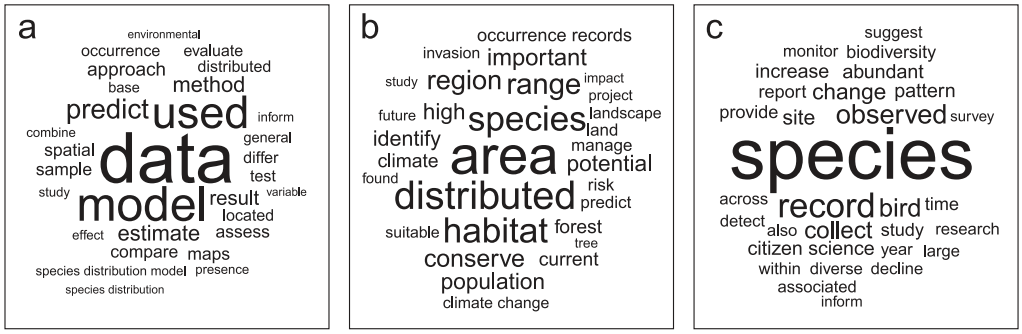


Figure 6: The articles were classified into three topic areas using latent dirichlet allocation (LDA) topic modeling, which uses word frequencies to cluster articles by topic. The 30 most heavily weighted words in (a) the methodological topic (n = 641), (b) the applied topic (n = 753), and (c) the records topic (n = 757) are shown in the present figure. Word size indicates relative weight within each topic.

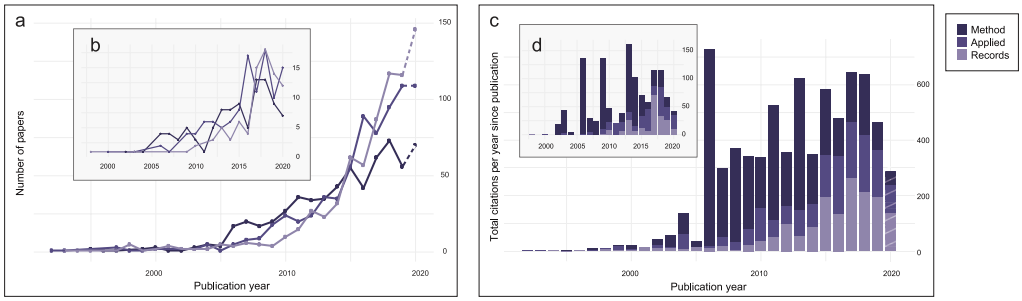


Figure 7: The number of articles published per year in each topic area within (a) the full set of 2151 articles and (b) the 300-article subset; the total citations per year since publication in each topic area within (c) the full set of 2151 articles and (d) the 300-article subset. 2020 is indicated with dashed lines because the results for 2020 may be less complete than those for other years; although the set of articles was obtained with a search on 4 January 2021, some articles with a 2020 publication date may not yet have been indexed by journals or the Web of Science.

recorded information on 10 fields: taxa, study system, study and author region, sample size, study scale, sampling design, analysis approach, data source, and data publication (see supplemental file S3). For all data fields except for study region and author region, the classifications were not mutually exclusive; each article was tagged with all applicable responses. Such classification is a common approach in descriptive literature reviews (e.g., Ball-Damerow *et al.* 2019, Hao *et al.* 2019). All data management and analyses were conducted with R version 4.0.2[52], and data and R scripts are available online[53].

Broad trends in the presence-only biodiversity literature

The literature relying on presence-only biodiversity occurrence data has grown steadily since the mid-2000s, maintaining an average annual growth rate that exceeds that of the biodiversity literature as a whole[54]. This literature has seen a shift in recent years from a focus on methodological research to data sharing and applied analyses, as is evidenced by both the number of articles published and the citations obtained by articles in each topic area (figure 7). The methodological topic area was most common from the mid-2000s through 2015. From 2015 to 2020, the frequency of articles within the methodological topic area remained relatively constant, whereas the frequency of applied and records articles increased rapidly. Methodological articles are overall the most highly cited, but the relative citation rate has declined since 2015 (figure 7). The shifting distribution of topic areas suggests that there are two distinct eras in the presence-only data literature: an era focused on methodological developments, which lasted from approximately 2005–2015 and an era with a greater focus on applications that began in 2015 and continues today. A similar trend has been reported among articles that rely on GBIF-mediated data[13].

The increase in articles focused on simple reports of occurrence is likely due to an increase in infrastructure and incentivization for data papers in recent years[5,18,55], and the parallel increase in applied research may indicate that presence-only approaches are being used more frequently to address issues of relevance to conservation and management[27,56,57]. The decline of methodological articles in terms of relative frequency and citation rate might suggest that applied researchers are using more established analysis methods more often than they are adopting newer approaches.

As a whole, the literature relying on presence-only biodiversity data is relatively decentralized and young. Its influence, as was measured by citations, is still growing; just a small number of the reviewed articles were highly cited, with a median of six citations per article. Unsurprisingly, methodological articles made up the majority of the 89 articles cited more than 100 times (figure 7; see supplemental file S5). The average author contributed to just 1.3 of the reviewed articles, which aligns with trends reported in the biodiversity literature[54] but is substantially lower than authorship rates in the biological sciences overall[58]. Articles were published in a wide range of outlets, with 482 distinct journals represented in our review. The relative lack of common references is a further indicator of the varied scope of the presence-only biodiversity literature (see supplemental file S5). This is likely due to specialization among biodiversity researchers within many distinct research areas, defined for example by taxon of interest, geographic region, or scientific subdiscipline. Nevertheless, it may indicate a challenge to the efficient sharing of information regarding best practices for biodiversity data sharing.

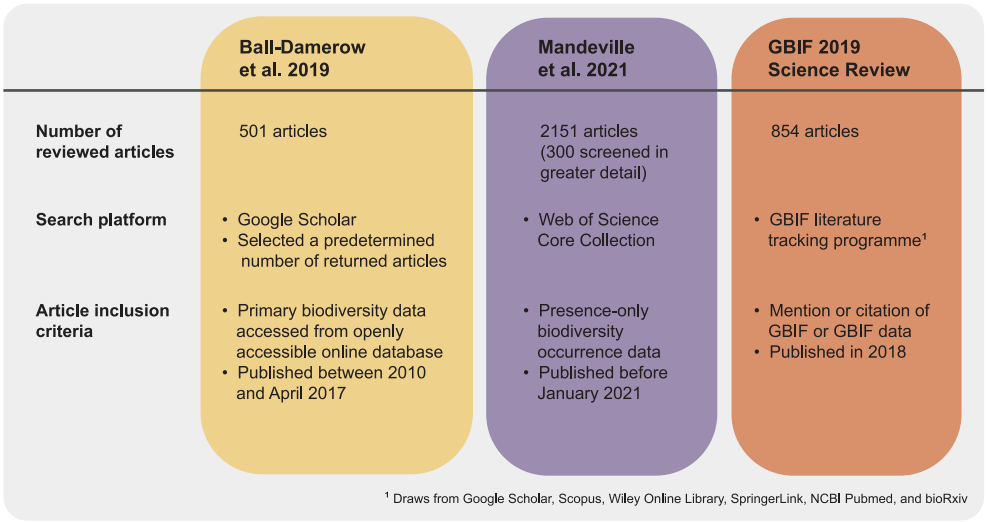
Using complementary reviews to build a more complete picture of the biodiversity literature

All efforts to systematically review literature contain trade-offs and biases introduced by the strategy used to search the literature, including search terms, search platform, and screening protocol. Therefore, efforts to characterize a body of literature are most informative when complementary reviews are considered alongside one another to form a more complete picture of the literature as a whole. We expect that this is particularly true for rapidly expanding research areas, including the presence-only biodiversity data literature; reviews of presence-only biodiversity data are complicated by the broad and rapidly developing variety of ways that this data type is accessed, analyzed, and referred to in the literature. To this end, we conducted a small test of the similarity of our search results to those of two recently published complementary reviews: Ball-Damerow *et al.* (2019 and the 2019 GBIF Science Review (GBIF 2019). Each of these reviews used a search strategy and platform that complements our own, targeting a distinct subset of the literature on applications of presence-only biodiversity data (figure 8).

For this test, we identified the articles from our review that met the inclusion criteria defined for each of the other two reviews, screened the abstracts of 50 articles randomly selected from each of the other reviews according to our own inclusion criteria, and identified the percentage of articles that were common to our review and each of the complementary reviews. There was relatively little overlap between the articles in our review and the other two reviews (figure 8). The lack of overlap illustrates the importance of considering complementary reviews alongside one another. Although other recent reviews, including the two considered in the present article, have focused largely on applications of presence-only biodiversity data known to be accessed from open sources, our review fills a key knowledge gap by characterizing a broad set of the traditional literature with an as yet unknown reliance on open databases.

Comparison of basic study characteristics with trends in biodiversity research

Our review joins several recent studies in identifying trends in basic characteristics of the biodiversity literature, including taxonomic focus, study domain, and study region[13,18,60]. We found that the articles in our review align some general trends in the biodiversity literature, including an emphasis on terrestrial settings[13,18,60] (figures 9 and 10). Still, there are some distinct trends associated with the articles in our review: vertebrates—and, to a lesser extent, invertebrates—are (better represented among our reviewed articles than in other reviews of the biodiversity literature, whereas plants and the freshwater domain are underrepresented[13,18,60] (figure 9)). The overrepresentation of vertebrates in our review is primarily due to



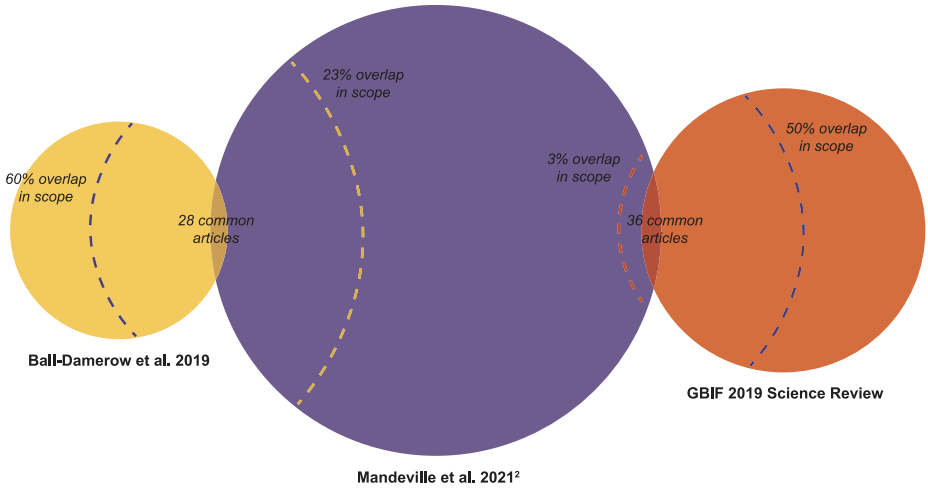


Figure 8: The Venn diagram indicates the overlap between articles included in this review and two complementary reviews. The circle size corresponds to review sample size; it should be noted that only a portion of the analyses reported in Mandeville (2021) were conducted on the full article set, whereas the remaining analyses were conducted on a subset of 300 samples chosen randomly from the full set. The overlap between the circles indicates the overlap in articles included in each review, and the dotted lines indicate the estimated overlap in targeted articles according to the reviews' described inclusion criteria. The inset table indicates the inclusion criteria and search strategy of each review.

their prevalence in reviewed articles that did not use data from open databases, suggesting that the range of vertebrate data available from open databases may not be as aligned with research needs as data from other taxonomic groups. On the other hand, the relative underrepresentation of freshwater and marine studies in our review was consistent between articles that did and did not rely on open data. This suggests that the presence-only data type as a whole may be less common in freshwater and marine domains, likely because many freshwater and marine species are not as easily detected via opportunistic observation.

The global distribution of studies in our review aligns closely with trends in the biodiversity literature[13,60]. The largest number of articles were authored by researchers based in Europe, followed by North America (figure 9). Alignment between study region and author region was uneven; articles that addressed Europe and North America were written by first authors based at institutions in the same region in respectively 98% and 95% of cases, whereas articles that addressed study regions in other parts of the world were less likely to have been written by first authors based in the focus region (figure 11). The uneven global distribution of biodiversity research reflects the greater coverage of biodiversity data in North America, Europe, and Australia relative to much of the rest of the world[15,61,62] and is also partially explained by the less frequent publication of ecological research conducted in the Global South in journals that are indexed by major databases[63]. It is critical that the field of biodiversity advances to better represent and support researchers based in underrepresented global regions in the international academic literature[63–65]. It has been shown that international collaborations are often inequitable, with European and North American researchers gaining more benefits in terms of publications and reputation than collaborators in the Global South[13,60,66–68]. This trend should prompt caution in the growing open data movement; it will be essential to ensure that open sharing of data is supportive rather than exploitative of Global South researchers[65,69–71]. One example of an approach to this issue from within the biodiversity data community is the ongoing effort to repatriate biodiversity data that have been collected within a historically exploited region but stored and managed elsewhere, in order to transfer primary data custody and decision-making power back to the communities from which the data were collected[13,70,72].

Presence-only data: A lens into current trends in the access, analysis, and publishing of openly accessible biodiversity data

As the biodiversity research literature continues to grow, the open sharing of biodiversity data is increasingly recognized as necessary and is quickly becoming normalized[13,17,18]. Presence-only biodiversity data are relatively representative of broad taxonomic and geographic trends associated with the field of biodiversity as

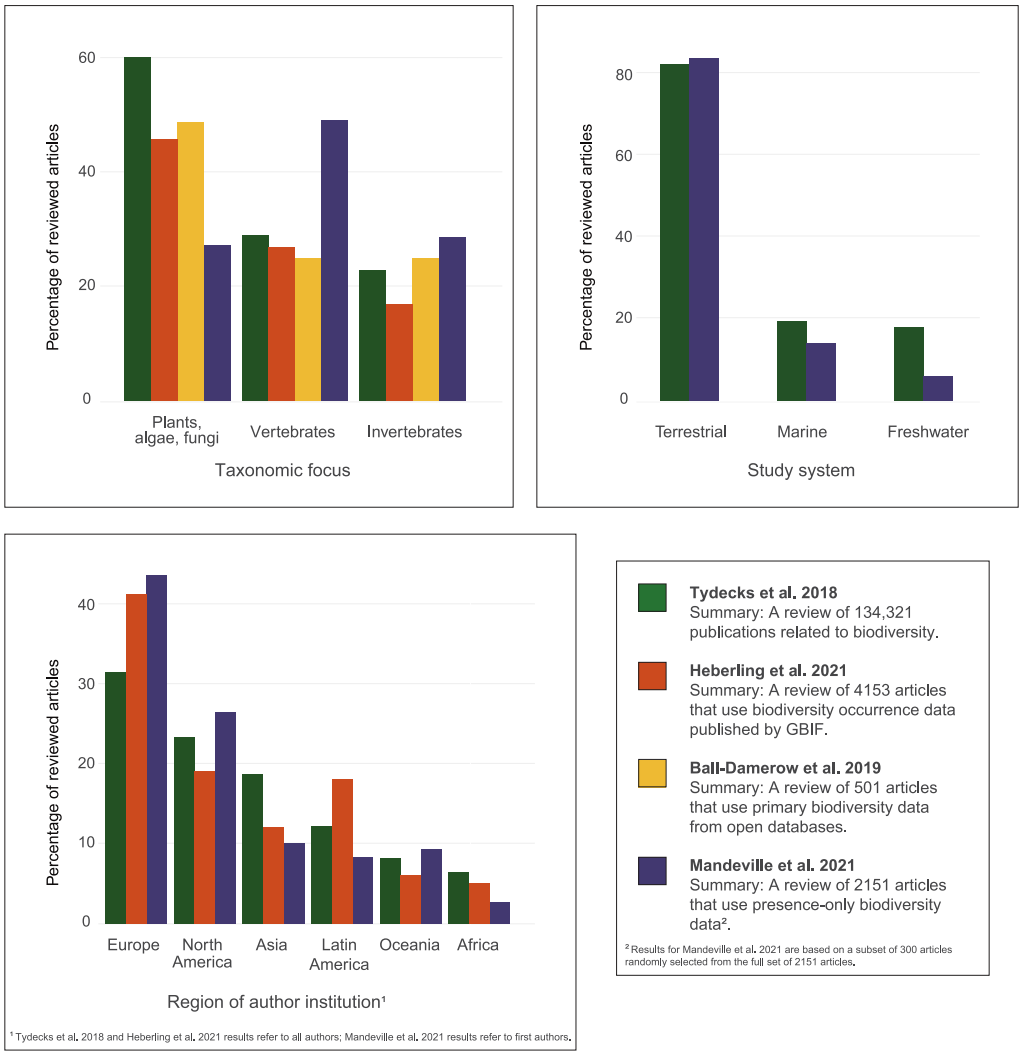


Figure 9: A comparison of trends in taxonomic focus, study system, and geographic region of the biodiversity literature identified by this review and three complementary reviews covering different aspects of the biodiversity literature. See each cited paper for specific methods and results, because the methods of defining and measuring each trend may differ slightly between articles.

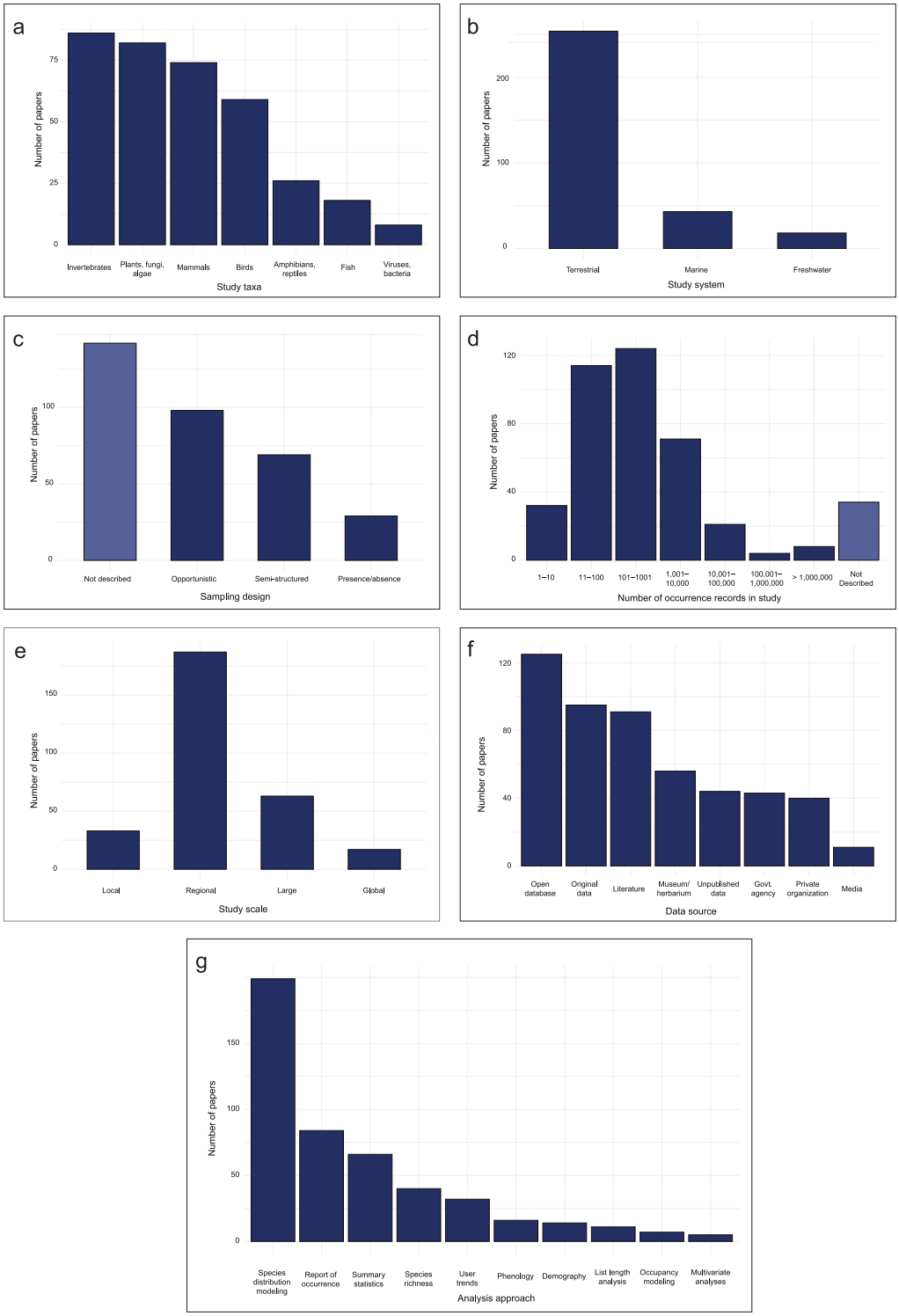


Figure 10: The frequency of characteristics among the subset of 300 randomly selected articles: (a) study taxa, (b) study system, (c) sampling design, (d) sample size, (e) study scale, (f) direct data source, and (g) analysis approach. Characteristics are not mutually exclusive; multiple responses per characteristic can apply to an article.

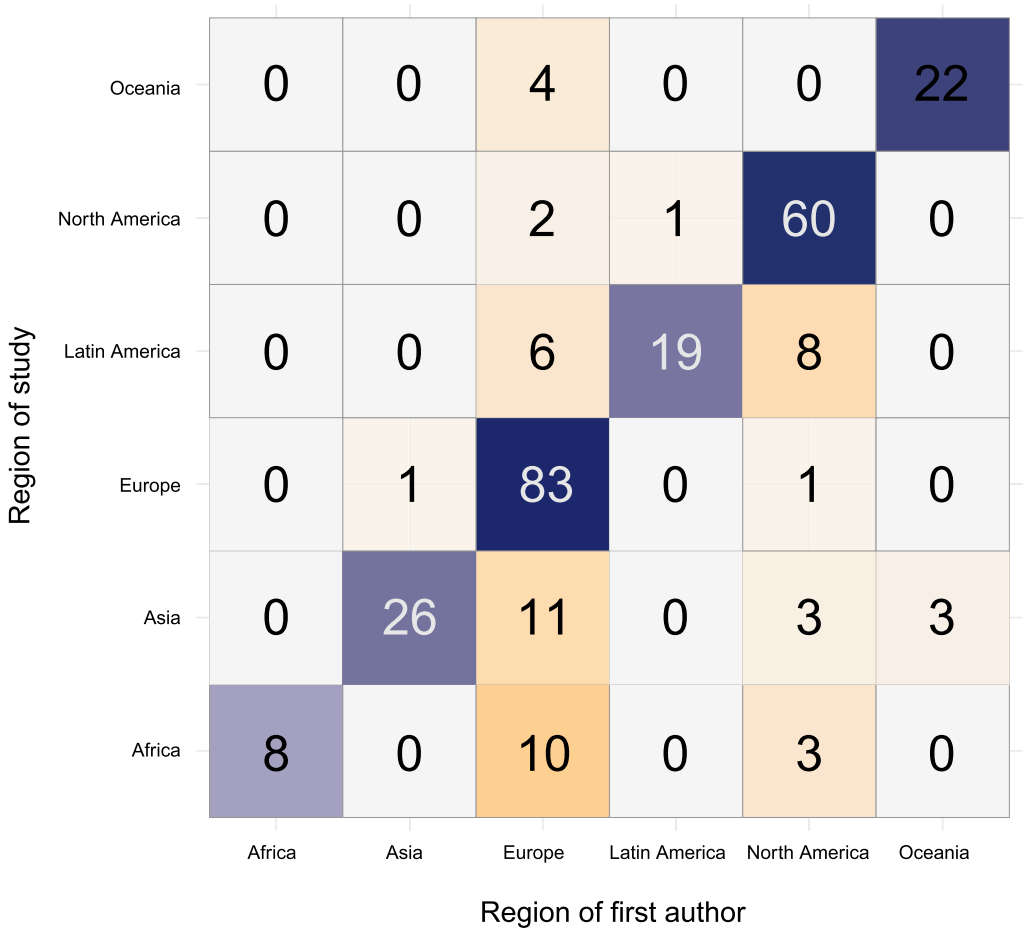


Figure 11: The study regions of the subset of 300 articles are indicated on the y-axis and the region of the first author of each article, defined by institutional affiliation, is indicated on the x-axis. The number in each cell indicates the number of articles written about the region on the y-axis by a first author based in the corresponding region on the x-axis.

a whole, but they differ in the ease with which they can be shared in accordance with currently recognized best practices[2,14–16]. Therefore, as practices continue to be developed to facilitate the sharing of a wide range of data types[2], presence-only data can serve as an early indicator to illustrate the progress, challenges, and limitations to the adoption of biodiversity data sharing practices. The work of recent reviews focused on presence-only data from open databases (e.g., Ball-Damerow *et al.* 2019 and the GBIF Science Review series) makes it clear that open data infrastructure actively supports a large body of research. But to understand the extent to which biodiversity research in the traditional peer-reviewed literature serves to facilitate or slow the progress toward open data, it is necessary to consider presence-only data from a wider range of sources.

In the sections that follow, we focus on three aspects of the presence-only biodiversity data literature indexed in the Web of Science Core Collection, with an emphasis on open data practices. We first consider the sources of presence-only data in this body of literature. Next, we consider how presence-only data are analyzed and whether these analyses are supported by well-documented metadata. Finally, we characterize the data publication practices associated with the presence-only biodiversity data in this set of literature. Our objective is to delineate the current state of data sharing practices and to identify areas for growth, many of which will apply to both presence-only data and also more generally to a range of biodiversity data types.

Sources of presence-only biodiversity data

Openly accessible databases—that is, searchable online repositories in which biodiversity data from many original sources are aggregated—make billions of biodiversity data points freely available for anyone to access and use[17,18]. Researchers may choose to access data from openly accessible databases for many reasons: to avoid duplicating research effort that has been undertaken in the past, to access data on a larger temporal and spatial scale than could be collected through original field work, to synthesize data from disparate sources, or to replicate or build on a previous study. So it is unsurprising that openly accessible databases were the most common direct data source in our review, accessed by 42% of the reviewed articles. However, only 19% of the reviewed articles used data exclusively from open databases; the vast majority accessed some or all of their data from sources other than open databases. Other common data sources include original fieldwork, the literature, and museums and herbaria (figure 10). Ball-Damerow *et al.* (2019 identified these same three sources of occurrence data as the most commonly integrated with occurrence data accessed from open databases.

In many cases, it is likely that researchers choose to collect new data or compile data from a variety of original sources because the data they need are not available in an openly accessible database[8,18]. For instance, articles in our review were

substantially more likely to address vertebrate species than in reviews in which all articles rely at least partially on open data (figure 9). In particular, a large percentage of the articles in our review addressed mammals (figure 10). Although mammals are considered overrepresented in open databases on a per-species basis, they make up a relatively small portion of the total volume of data available from open databases, likely because of many mammal species' lower detection probability, wider-ranging distributions, and relatively lower dedicated citizen science interest than some other taxa[8,73]. This may explain why articles that addressed mammal species were relatively unlikely to obtain data from an open database and more likely to obtain data from government agencies, private organizations, and through original data collection. Overall, the relatively small percentage of articles based on open presence-only data corroborates a growing sentiment from the literature: Although the volume of openly accessible biodiversity data continues to grow, there are substantial taxonomic and spatial gaps for which there is minimal open data[8,18,74–78]. Our results corroborate the many studies that have identified gaps in biodiversity data, making it clear that the majority of researchers who conduct presence-only analyses do not find the data they need in open databases. This highlights the need for the biodiversity research community to continue ongoing efforts to identify and fill critical taxonomic and spatial knowledge gaps in open databases.

Data gaps can be filled through both novel data collection and mobilization of existing data that are not yet openly accessible. Many large pools of data exist outside the open data infrastructure—for example, in government agencies and private organizations[77,79,80]. Identifying these sources of data, supporting policies and infrastructure that facilitate their access and reuse, and incentivizing data sharing at an institutional level is needed to facilitate more open access to these data[81]. This is critical for establishing the long-term records that are essential for studying trends across space and time and informing conservation interventions in the face of global change[77]. Opening existing data for reuse is also necessary to avoid duplication of data collection effort and research waste, freeing research resources to target true data gaps[82]. Consider, for example, that 13% of the articles in our review accessed data from 10 or more nonopen sources, some accessing well over one thousand distinct sources. The collation of data from multiple sources represents an extensive research effort that will likely need to be repeated by future researchers if the data are not made more openly accessible. Reducing inefficiencies by supporting the access and reuse of data will allow researchers to prioritize generation of data that will fill gaps in the available knowledge. To achieve this, efforts to build relationships between data aggregators and the research community will continue to be essential.

In other cases, openly accessible data may be available to replace or supplement data from other sources but authors may neglect to use it, either because they are not aware of it or because they do not trust its quality[83]. Even when data are aggregated in an open database, some researchers may choose to access the

data from their original sources rather than from the open database[84]. In some cases, researchers may be aware of open data but believe they lack the skills to access and use it effectively[43]. Indeed, a broad survey of researchers found that the perceived value and efficiency of reusing open data were major factors in whether researchers chose to access open data[85]. Finally, it is also important to note that inequities in technological infrastructure, competence, and training mean that access to digital platforms is also inequitable[86]. Finding solutions to the barriers that keep researchers from accessing open biodiversity data should be a goal of the biodiversity research community.

Practices for accessing and citing open data vary widely

Among open databases, data sources varied widely. We identified 117 open databases that were used to access presence-only occurrence data (see supplemental file S6). We classified nine of these as large open databases, defined as relatively well known, established databases that contain data covering a very large geographic range, a wide range of taxa, or both. The most commonly accessed was the GBIF, which was accessed by 37 articles, followed by eBird (9 articles). The remaining 108 open databases, classified as small databases, had a narrower geographic or disciplinary scope and were each accessed by an average of 1.2 articles. Of the articles that accessed open data from at least one source, 55% accessed a large database and 65% accessed a small database. Two thirds directly accessed just one database, whereas the remaining third accessed between two and 10 distinct open databases. Of course, because many open data sources serve to aggregate many smaller databases, data users that accessed just one database may still have obtained data from a wide range of original sources. These results are similar to the findings of Ball-Damerow *et al.* (2019), who also found that a small number of open data sources were cited by many articles, whereas a large number of open data sources were cited very few times.

The frequent reliance on small open databases is probably due in large part to the prevalence of small databases within specific research areas[18,84,87] and may also be partially explained by a lack of familiarity with or trust in large databases[83]. We recognize many values of small databases, including responsiveness to specific disciplinary requirements[88] and the cultivation of strong relationships between data curators and communities of data users[89,90]. However, small open databases may lack the standardization and interoperability that are built into larger data aggregators[43], they may lack consistent leadership to maintain growing content and keep up with developing best practices[6], and they are more likely to become technologically obsolete, rendering the data inaccessible[18,35,89,91].

We attempted to access all of the databases referred to in our reviewed articles and found that we could not locate or access 9% of the small databases from which articles in our review had obtained data. In a few other cases, the database

website could be accessed, but it was not clear that the data were still accessible; for example, data could be visualized but the link to download data was broken, or it was requested that visitors contact the database managers to request access. Although still concerning, it is perhaps a cause for cautious optimism that the proportion of inaccessible databases in our review is considerably lower than the 26% of databases found to be inaccessible by Ball-Damerow *et al.* (2019), who reviewed articles published through April 2017. An additional 15% of the small databases had been consolidated into a different database but were still accessible. All nine large databases remained accessible. Because of the important role played by small databases, we do not intend to suggest that authors avoid them; rather, we caution the biodiversity data community to be cognizant that these small databases are strongly relied on and to be proactive about supporting them over time[87]. The true reliance on small databases is likely to be even higher than identified in our study because small regional databases may be cited more frequently by articles published in regional journals and gray literature, which may not be indexed by the Web of Science and so may have been underrepresented in our search[92].

The proliferation of open data aggregators, along with the rapidly evolving best practices for their use, has resulted in an uneven landscape of how such data are cited in the literature[18,45,46]. Citation of a digital object identifier (DOI) that is uniquely connected to the full data set analyzed in an article has emerged as the best practice in this area[7,13]; this practice enables the data set to be clearly replicated and all original sources to be credited[45,46]. But not all researchers are yet aware of this best practice, because it is relatively new. Furthermore, not all open databases have a clear mechanism for producing a citable DOI[93,94]. We found a great deal of variation in how open databases were cited among the articles in our review. The vast majority of articles simply listed the names of the databases from which they obtained data, sometimes accompanied by a brief description of the type of original sources from which the data were aggregated. Only 4% of the data sets accessed from an open database were cited with a DOI, and another 3% were not cited but, instead, were described in the text of the article with a direct link to the full data set or other thorough directions that would enable a reader to replicate the data retrieval process. Interestingly, the proportion of articles in our review that included a database citation with a URL or DOI was much lower than the 34% observed by Ball-Damerow *et al.* (2019). This may reflect a difference in search strategy; the search terms used by Ball-Damerow *et al.* (2019) ensured that all reviewed articles at least mentioned the type of database accessed, whereas our search terms required only that articles mentioned the type of data. The differing results obtained by these two searches suggest that the use of appropriate citation practices may be correlated with authors' use of specific terminology to refer to open databases, perhaps signaling their perception of their work as related to the open data movement.

A small number of authors in our review found alternative ways to recognize original providers of data even when there was no mechanism to do so through the

open database—for example, by listing all original data sources in the supplemental material. Giving credit to the original providers of open data is critical for incentivizing data sharing to researchers, institutions, and funders[18,45,95] and for recognizing and supporting the diverse landscape of organizations and institutions that engage in biodiversity monitoring[9]. This may be especially true when data were collected through public involvement in citizen science. Thirty-four percent of the articles in our review identified citizen science as the original source of some or all of their data, although the true percentage of articles that derived data from citizen science is likely higher because citizen science data are frequently reused without their source being clearly described[96]. Citizen science plays an important role in biodiversity data collection but long-term funding and support for many citizen science programs may be dependent on the demonstrated impact, so appropriate citation is critical[97–100].

Analysis and reporting of presence-only biodiversity data and associated metadata

The growth of interest in presence-only data in the mid-2000s was paralleled by innovation in species distribution modeling approaches tailored to this data type[18,38,101], so it is unsurprising that species distribution modeling was the dominant analysis approach in our review (figure 10). These methods have become increasingly sophisticated and widely popular[51,102,103]. A large review of articles that use GBIF data found a similar prevalence of species distribution modeling and identified a recent transition in focus from methodological developments to widespread application similar to that seen in our overall set of reviewed articles[13]. Although the initial development of species distribution modeling approaches for presence-only data was at least partially a response to the increased availability of the data type, we suggest that their subsequent wide adoption has created a positive feedback effect whereby researchers, driven by the growing ease of analyzing presence-only data, have increasingly begun to seek out presence-only data from a wider range of sources.

Despite its prevalence, however, species distribution modeling is far from the only analysis method applicable to presence-only data. Our results illustrate a wide range of analysis approaches, including both inferential statistics and a variety of descriptive statistics. Presence-only data are also occasionally used indirectly—for example, to validate the results of another analysis or to inform a sampling design. Methodological innovation in inferential approaches is ongoing, and since 2012, a number of articles have applied a variety of less common inferential approaches, including phenology analyses, demography analyses, list length analysis, occupancy modeling, and multivariate statistics (figure 10). In particular, the integration of presence-only data with other types of biodiversity data is of growing interest in the literature[104–109]. In our review, articles that integrated presence-only data

with other types of biodiversity data were nearly three times as likely to employ an uncommon inferential analysis approach as the articles that used only presence-only data, indicating that data integration can open a wider range of analysis options for presence-only data.

Clearly documented metadata, particularly an explicit description of the data structure and original sampling design, also enable a wider range of analytical approaches, including data integration[38,41,110]. This trend is reflected in our results, with articles that employed more complex analysis approaches being correspondingly more likely to describe the underlying data structure (figure 12). Articles that employ species distribution modeling are the major exception to this trend; despite the relative statistical complexity of species distribution modeling, articles that modeled species distributions were the least likely to document data structure (figure 12). This likely reflects the growing accessibility of species distribution modeling approaches, which have become increasingly straightforward to implement through user-friendly platforms that can be implemented as a black box by researchers without a clear understanding of the method[111–113]. Although the growing accessibility of species distribution modeling offers great potential for research and conservation[114,115], we caution that it is still essential to share metadata whenever possible to aid in interpretation and evaluation of results[42,103,116–118]. Relatedly, it is important to check for and correct data quality errors in data and metadata, particularly when data are obtained from open databases or collated from several sources[18]. In addition to supporting data interpretation and analysis, the reporting of high quality metadata facilitates a wide range of potential future data uses.

Reporting of metadata is inconsistent

Despite the value of clear metadata, around half of the articles that we reviewed did not explicitly describe the structure or sampling design of all of their data, corroborating previously reported trends[119,120] (figure 10). Of course, researchers can only report metadata if they have access to this information, and researchers reusing data may simply not have information on the original data structure. For instance, 118 articles obtained data from museums, herbaria, and the literature and 77% of these did not report the structure of their data; in the vast majority of these cases, metadata on the original sampling design were likely unavailable. Users of open data also have inconsistent access to metadata, and around half of the articles that obtained data exclusively from open sources did not describe data structure (figure 12). Although many openly accessible databases enable and encourage metadata standardization and sharing, most prominently through the Darwin Core standard[12], many data available through open databases have been digitized from historical records, for which such metadata may be unavailable or may have been lost over time[121]. Articles that rely on data collected by government

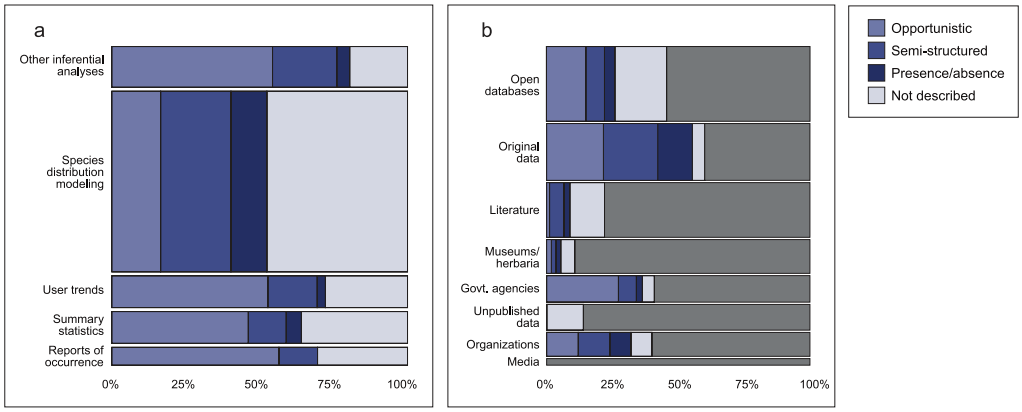


Figure 12: The percentage of the 300-article subset that is associated with each type of data structure, as a function of (a) analysis approach and (b) direct data source accessed by study authors. In panel (a), the y-axis categories represent all articles for which the indicated analysis approach was the most complex approach applied (with the exception of "user trends," in which case all articles using this approach are represented). The bar widths indicate the number of articles in the 300-article subset within each category. In panel (b), the y-axis categories represent all articles that use data from the indicated data source. The bar widths indicate the overall proportion of the 300-article subset that used each data type. The gray portions of the bars represent articles that integrated data from the indicated source with data from other sources; because of the confounding effect of data integration on metadata reporting, metadata reporting trends are not reported for these articles. The portions of the bars shaded according to the legend represent articles for which the indicated source was the only source accessed by the article.

agencies and private organizations describe data structure more frequently (figure 12). In the instances in which the structure of data from these sources is not described, it may be due to the loss of information that occurs when complete information was not passed from the data owners to the data users. Standardizing the methods used by governmental and private institutions to share data with researchers may reduce instances of data loss associated with more informal sharing of data[9]. Unsurprisingly, articles exclusively based on original field work were most consistent in documenting data structure (figure 12). The combination of data from multiple sources is an additional barrier to describing presence-only data because of practical challenges associated with describing a large number of separate sampling schemes. For each additional source accessed by an article in our review, the likelihood of data structure being described decreased by 12%. Although authors may have little recourse when working with data sets for which metadata are unavailable or with large data sets for which it may be impractical to describe a large number of separate sampling schemes, improving data citation practices may provide a partial solution by making it possible to trace data to its original source to gather any available metadata.

Of articles that described the structure of their data, most described one or more data source as opportunistic (i.e., collected with no predefined sampling design), followed by semistructured (*sensu* Dobson *et al.* 2020), and finally a smaller percentage used presence or absence data and discarded the absence records before analysis. Of the articles that converted presence or absence data to presence-only format before analysis, one third did this for the purpose of comparing different modeling approaches. The remaining two thirds discarded the absence data and conducted analyses exclusively in a presence-only framework. Previous authors have cautioned that it is not advisable to analyze presence or absence data in a presence-only framework[122], so it is concerning that some articles in our review took this approach. In some cases researchers may be motivated to convert presence or absence data to presence-only to facilitate merging presence or absence and presence-only data sets, but many recent studies suggest approaches for integrating various data types without reducing data structure[104–107,109].

The articles in our review were more consistent in reporting the scope of their presence-only data set, in terms of both sample size and study scale. The sample size varied considerably between articles, but the majority of studies were small to mid-size (figure 10). The studies' geographic scale followed a similar trend, with the majority addressing a regional scale (figure 10). The small number of articles that did not explicitly state a sample size tended to involve several separate analyses of a large number of species and stated a total sample size and total number of species rather than the sample size for each analysis. The tendency toward mid-size studies has remained relatively consistent over time, with the exception of studies with a sample size of over one hundred thousand occurrence records. These very large studies were absent from our reviewed articles until 2014. This recent increase in large studies likely reflects growing infrastructure for and

interest in big data macroecology[15,123]. Such large studies are more likely to rely on open data than studies with a smaller scope.

How often are presence-only data made available for reuse?

Our results suggest that the majority of data used in presence-only analyses are not made available after the analyses are published, although there is a recent trend toward increased data sharing. To characterize trends in data sharing, we excluded the 19% of articles that were based entirely on data accessed from open sources. Of the remaining articles that used data from at least one source other than an open database, just 21% made all data used in the study openly available on publication of the article. Of these, 18% published their data in an openly accessible online database, whereas the rest used a different form of publication, such as supplementary material or an online repository (figure 13). The most common means of sharing data was to directly include it in the article, either the main text or the supplemental material. Data formats varied from those that facilitate reuse relatively easily (e.g., CSV files, spatial data files) to those that pose challenges for reuse (e.g., PDF files). Online repositories, including Dryad, Figshare, and GitHub, were also used by a small number of articles to share data. Only nine articles indicated that their data sets had been shared in an openly accessible database, although it is possible that the authors of some articles in our review published their data to an open database but neglected to mention this in the article. Of course, the data analyzed in the 19% of reviewed articles that obtained data exclusively from open databases remained openly available as long as the databases from which the authors accessed their data were still accessible.

To maximize their research value, data must be published in a way that is both searchable and persistent[10,44]. Therefore, publication of data in aggregated databases is preferable to publication in supplemental material. In particular, larger databases are more likely to have greater longevity, stability, and infrastructure to maintain current best practices for data management in this rapidly developing field[43,87]. Much like small open databases, it has been demonstrated that data in supplementary material often become inaccessible over time[91,124]. We attempted to access all data shared by our reviewed articles and found that it was largely, but not entirely, still accessible: 7% of the data sets shared in journal supplementary materials were no longer available, and 22% of the data sets shared in an open database were no longer available. The inaccessible data from open databases were exclusively shared in small databases.

Although the overall accessibility of openly available presence-only data has increased dramatically in recent years, our results make it clear that the traditional peer-reviewed literature still largely serves as a sink for presence-only biodiversity data rather than facilitating its sharing and reuse. Making presence-only data

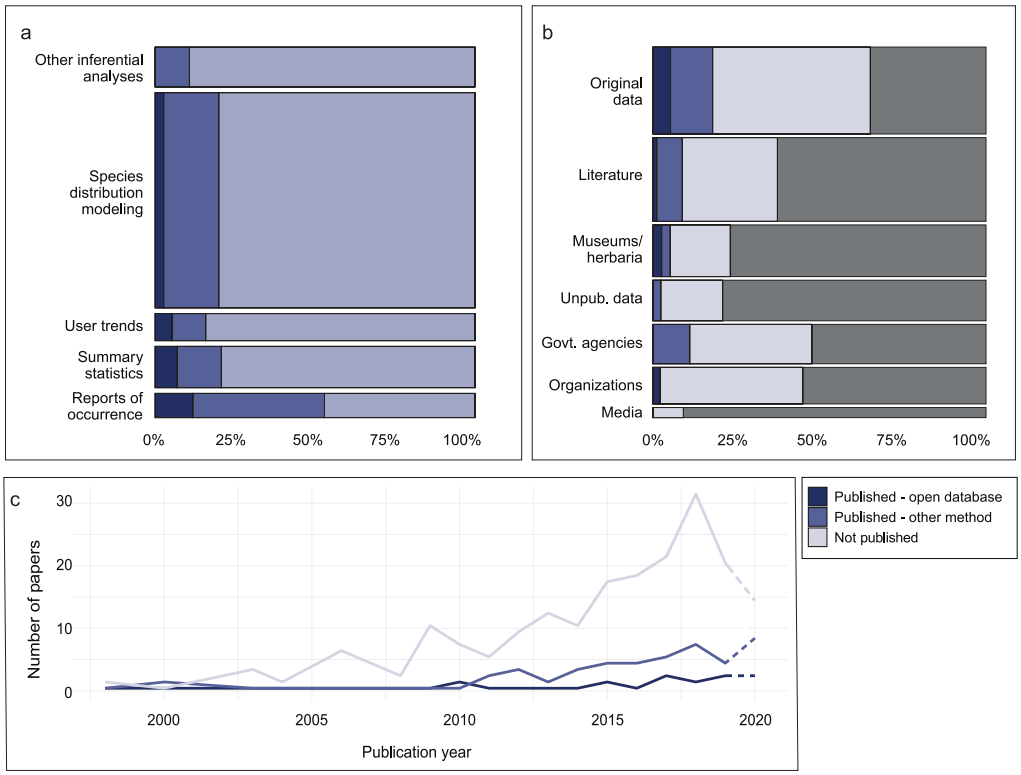


Figure 13: The percentage of the 300-article subset that is associated with the three levels of data availability as a function of (a) analysis approach and (b) direct data source accessed by study authors. For all panels of this figure, articles based entirely on data accessed from open databases have been excluded, leaving a subset of 242 articles that access data from at least one source other than an open database. In panel (a), the y-axis categories represent all articles for which the indicated analysis approach was the most complex approach applied (with the exception of "user trends," in which case all articles using this approach are represented). The bar widths indicate the total number of articles within each category. In panel (b), the y-axis categories represent all articles in which the indicated direct data source was accessed. The bar widths indicate the overall proportion of the 242-article subset that used each data type. The portions of the bars shaded according to the legend represent articles for which the indicated source was the only source accessed by the article or which integrated the indicated source with open data. The gray portions of the bars represent articles that integrated data from the indicated source with data from other sources; because of the confounding effect of data integration on data sharing, data sharing trends are not reported for these articles. Panel (c) indicates trends in data availability over time. 2020 is indicated with dashed lines because the results for 2020 may be less complete than those for other years; although the set of articles was obtained with a search on 4 January 2021, some articles with a 2020 publication date may not yet have been indexed by journals or the Web of Science.

more accessible should be a clear priority. Because strong infrastructure and clear best practices already exist for sharing presence-only occurrence data[2,17,87,125] this should be achievable. However, several barriers can stand in the way of data sharing, including researchers' lack of incentive and ability, data ownership, and data set complexity. The strategies for overcoming these barriers will differ on the basis of the original source, ownership, and structure of the data.

Data sharing considerations for different types of presence-only data

The most straightforward type of presence-only data to target for increased data sharing are likely those collected by the study authors. Our results do indicate that original data are the most frequently shared, but the sharing rate is still low, at just 27% (figure 13). The publishing rate of original data collected with citizen science was somewhat higher than average, although still fewer than half of the articles based on original citizen science published their data. This is problematic, because studies have shown that citizen science participants generally expect and want their data to be made available for research, conservation, and policymaking[97,126–129]. Further integration of citizen science with open biodiversity data aggregators should therefore be a priority.

We anticipated lower rates of data publication from articles that compiled data from third party data owners, including the literature and museums and herbaria, and our results indicated rates of data publication that were just slightly lower than that of original data (figure 13). We suggest two major reasons why authors may not share data they have collated from other data owners. First, they may lack (or perceive that they lack) the permission to do so. And second, they may perceive that data sharing is unnecessary, assuming that readers wishing to reproduce their data set could retrace the data acquisition methods described in the paper to reassemble the data set from its original sources. Although this may sometimes be true, collating data from multiple sources takes a great deal of time and effort, so it is not a trivial process for a reader to reassemble a data set following a process described in the literature. And even if original data sources are well documented and still accessible, it cannot be assumed that a reader will be able to replicate the steps taken to collect data; literature is often behind paywalls, and access to institutional databases may be limited. Therefore, researchers working with data compiled from museums, herbaria, and journal articles should strive to provide as thorough a description as possible of their exact process of compiling their data set or, better yet, publish their complete data set whenever possible[130]. Widespread progress on this issue will depend in part on the support of institutions: Institutions that host data should institute mechanisms to generate citations when data are accessed, making data easier to cite[131–133], and journals that publish research should outline clear policies that support and facilitate data sharing and

citation[134].

Finally, there are circumstances in which researchers may be unable to share data because of its proprietary or sensitive nature. We expect that this issue is most relevant to data obtained from private organizations or government agencies; in the present review, articles that accessed data primarily from one of these sources were characterized by low rates of data publication (figure 13). This is a complex issue, but we would encourage owners of sensitive data to use existing decision tools and prioritization schemes to consider whether there is a suitable way to make these data available for reuse, even in a more limited format[57,135,136]. Because 37% of reviewed articles derive at least a portion of their data from sources that are assumed to generally be nonopen (e.g., data provided by government agencies, private organizations, or personal communications), and 41% derive some or all of their data from sources that are potentially accessible but cannot be assumed to be available to all readers (e.g., museums, literature, media), it is clear that a large portion of the presence-only biodiversity literature relies on data that are not accessible, hampering the replicability of these studies and the reusability of the data on which they are based.

A separate but related issue concerns data ethics and ownership. Issues of data ownership and governance are inherently related to social governance, and it is essential that the ethics of data sharing be held in the forefront at all stages of data management[71,137,138]. Data relevant to local communities must be made accessible to community members and must not be used in ways that are counter to community priorities[86]. This is particularly essential when it comes to Indigenous data; the CARE Principles for Indigenous Data Governance are a critical framework for ensuring Indigenous peoples' rights to the control of Indigenous data[137,139]. In addition, when data are collected by community members, as with citizen science, it is important to understand and respect volunteers' motivations for and concerns about the use of data they have contributed[126,140,141]. The continued normalization of open data sharing must center scholarship and practice that respects ethical data governance, stewardship, and access.

The future of presence-only biodiversity data sharing

Data sharing practices in the presence-only biodiversity literature have until recently remained relatively constant over time, but the proportion of reviewed articles that publish their data has increased somewhat since 2016 (figure 13). This is cause for optimism and continued efforts to normalize open sharing of biodiversity data. Recent studies document overwhelmingly positive attitudes to data sharing[142,143], so if practical barriers can be overcome, there is a high likelihood that data sharing will continue to increase. Increased sharing of biodiversity data may even produce a ripple effect across disciplines; biodiversity research has historically exhibited a higher rate of open data sharing than closely

related scientific disciplines such as ecology and conservation science[144–146], but given the broad and growing application of presence-only biodiversity data across many related scientific disciplines[13,18], continued improvements in open sharing of presence-only biodiversity data may serve to spread awareness of open data practices across disciplines.

Past studies have indicated that the majority of biodiversity researchers support data sharing but may be held back by lack of sufficient incentive, lack of familiarity with data aggregators, lack of information on data set structure or ownership, and lack of trust in public databases[142,147]. We compared articles that did and did not publish their data to examine the relative impact of some potential barriers to data sharing. First, we anticipated that two measures of data set complexity might negatively correlate with data sharing: first, the number of data sources accessed to compile a data set and, second, whether the original sampling design was reported. We expected that authors might be held back from sharing data by the complexity of crediting multiple original sources or by their own lack of complete information on data structure. However, we did not find either of these relationships in our results. This finding suggests that data set complexity may not be the primary factor prohibiting researchers from publishing their data sets. It is a concern but is more likely secondary to other barriers. Because lack of familiarity with open databases has also been cited as a barrier to data sharing, we expected that authors' familiarity with open data, as has been demonstrated by the integration of data from open databases with presence-only data from other sources, would correlate with greater rates of data publication. This was not the case: Of the articles that integrated data from open databases and other sources, 76% did not publish the data that were not already open.

These findings suggest that other concerns, including lack of researcher incentive and concern about receiving appropriate credit for shared data, may be more serious barriers to data sharing[45,142]. Some developments have begun to address the issue of researcher incentive: Data sharing is increasingly incentivized through journal policies, funding agency requirements, and the promotion of data citations[148–150]. Continuing to normalize these incentives may help overcome existing barriers to data sharing, especially in situations in which data users are the original data owners[45,131,151,152]. Furthermore, researchers are increasingly taking ownership over the process of data sharing, establishing grassroots collaborations that organize specific research communities to engage with open data infrastructure and practices[153]. This integration of open data practices into local networks of biodiversity researchers has great potential to incentivize open data sharing by establishing it as a key component of network building and collaboration within specific research areas. As open data sharing becomes increasingly normalized, it will be essential that practitioners of open science maintain a supportive, rather than critical, approach to encouraging researchers who are taking their first steps into open data sharing. Researchers do not all have equal access to the resources, training, technical capacity, and institutional support to fully engage in open data

practices, and small steps toward open data sharing must be welcomed while the field as a whole shifts to become more equitably supportive of open data practices[5,133,143,154].

Conclusions

Open access to high quality biodiversity occurrence data is key to many emerging themes in biodiversity research and conservation, including development and implementation of international biodiversity assessments and targets[78], research synthesis for conservation decision-making[155], and near-term ecological forecasting of species abundance in space and time[156], so continued efforts to increase the open sharing of biodiversity data will be critical. This will require increased incentivization, institutional support, ongoing shifts in cultural norms, and a growing emphasis on an ethical, equitable framework for data sharing. Recent trends toward increased sharing of presence-only biodiversity data are a cause for optimism, but there is still a great deal of work to be done in normalizing the use of best practices in data access, documentation, citation, and sharing. Still, we see evidence in the trends reported in the present article for an often-reported survey result: Researchers generally feel positively toward reusing and sharing data, despite persistent uncertainty about best practices and concern about credit and incentives[142,143,157]. Such evidence includes the recent increase in the proportion of articles that produce open data, the efforts taken by some authors to credit original data providers even when no clear mechanism had yet been developed to do so, and the above-average sharing rate for citizen science data.

For researchers looking to begin or continue their journey into reuse and sharing of open biodiversity data, there are many excellent resources that offer an entry point into accessing and sharing open data; we particularly point such researchers to Hampton *et al.* (2015), Wilkinson *et al.* (2016), Boland *et al.* (2017), Alston & Rick (2021), and to guides such as the FAIR Principles[160], the CARE Principles of Indigenous Data Governance[139], and the Quick Guide to Publishing Data Through GBIF.org[gbif2021quick]. To those beginning to engage with open data, we echo the wisdom of Bahlai *et al.* (2019), Alston & Rick (2021), and others in encouraging researchers to begin with any first steps, however small, that are feasible given their circumstances. Increased open data sharing will rely on both the progressive adoption of data sharing practices by individual researchers and ultimately on broad cultural shifts within biodiversity and related fields[5]. This shift to a culture of ethical open data sharing will be essential to meet challenges associated with the growing biodiversity crisis and to support a growing need for biodiversity assessment, monitoring, and conservation.

Acknowledgments

This work is part of the Norwegian University of Science and Technology's Transforming Citizen Science for Biodiversity project, and we thank the other project members for their valuable feedback on earlier stages of this work. We also greatly appreciate the feedback of three anonymous reviewers and an anonymous editor. Funding for CPM was provided by the Transforming Citizen Science for Biodiversity program within the Digital Transformation Initiative of the Norwegian University of Science and Technology. WK was funded by the Norwegian Research Council (grant no. 272947) and the Norwegian Biodiversity Information Centre. EBN was funded by the Norwegian Institute for Nature Research.

Author Biographical

Caitlin P. Mandeville (caitlin.mandeville@ntnu.no) is a PhD candidate in the Centre for Biodiversity Dynamics and Department of Natural History at the Norwegian University of Science and Technology, in Trondheim, Norway. Wouter Koch is a senior advisor of biodiversity informatics at the Norwegian Biodiversity Information Centre, as well as a PhD candidate in the Centre for Biodiversity Dynamics and Department of Natural History at the Norwegian University of Science and Technology, in Trondheim, Norway. Erlend B. Nilsen is a senior research scientist at the Norwegian Institute for Nature Research in Trondheim, Norway, as well as a professor of ecology at the Faculty of Biosciences and Aquaculture at Nord University, in Steinkjer, Norway. Anders G. Finstad is a professor in the Centre for Biodiversity Dynamics and Department of Natural History at the Norwegian University of Science and Technology, in Trondheim, Norway.

References

1. Farley, S. S., Dawson, A., Goring, S. J. & Williams, J. W. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience* **68,** 563–576. `https://doi.org/10.1093/biosci/biy068` (July 2018).
2. Anderson, R. P. *et al.* Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Frontiers of Biogeography* **12.** `https://doi.org/10.21425/F5FBG47839` (2020).
3. Kays, R., McShea, W. J. & Wikelski, M. Born-digital biodiversity data: Millions and billions. *Diversity and Distributions* **26,** 644–648. `https://doi.org/10.1111/ddi.12993` (2020).

4. Alston, J. M. & Rick, J. A. A beginner's guide to Conducting reproducible research. *Bulletin of the Ecological Society of America* **102,** 1–14. https://doi.org/10.1002/bes2.1801 (2021).

5. Chawinga, W. D. & Zinn, S. Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research* **41,** 109–122. https://doi.org/10.1016/j.lisr.2019.04.004 (2019).

6. Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q. & Bourne, P. E. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution* **28,** 454–461. https://doi.org/10.1016/j.tree.2013.05.002 (2013).

7. Brown, R. F. The importance of data citation. *BioScience* **71,** 211–211. https://doi.org/10.1093/biosci/biab012 (2021).

8. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7.** https://doi.org/10.1038/s41598-017-09084-6 (Aug. 2017).

9. Kühl, H. S. *et al.* Effective biodiversity monitoring needs a culture of integration. *One Earth* **3,** 462–474. https://doi.org/10.1016/j.oneear.2020.09.010 (2020).

10. Wilkinson, M. D. *et al. The FAIR Guiding Principles for scientific data management and stewardship* Mar. 2016. https://doi.org/10.1038/sdata.2016.18.

11. Robertson, T. *et al.* The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PloS one* **9,** e102623. https://doi.org/10.1371/journal.pone.0102623 (2014).

12. Wieczorek, J. *et al.* Darwin Core: an evolving community-developed biodiversity data standard. *PloS one* **7,** e29715. https://doi.org/10.1371/journal.pone.0029715 (2012).

13. Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B. & Schigel, D. Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences* **118.** https://doi.org/10.1073/pnas.2018093118 (2021).

14. König, C. *et al.* Biodiversity data integration—the significance of data resolution and domain. *PLoS biology* **17,** e3000183. https://doi.org/10.1371/journal.pbio.3000183 (2019).

15. Wüest, R. O. *et al.* Macroecology in the age of Big Data–Where to go from here? *Journal of Biogeography* **47,** 1–12. https://doi.org/10.1111/jbi.13633 (2020).

16. Gadelha Jr, L. M. *et al.* A survey of biodiversity informatics: Concepts, practices, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **11,** e1394. https://doi.org/10.1002/widm.1394 (2021).

17. Peterson, A. T., Asase, A., Canhos, D. A. L., de Souza, S. & Wieczorek, J. Data leakage and loss in biodiversity informatics. *Biodiversity Data Journal.* https://doi.org/10.3897/BDJ.6.e26826 (2018).

18. Ball-Damerow, J. E. *et al.* Research applications of primary biodiversity databases in the digital age. *PloS one* **14,** e0215794. https://doi.org/10.1371/journal.pone.0215794 (2019).

19. Theobald, E. J. *et al.* Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation* **181,** 236–244. https://doi.org/10.1016/j.biocon.2014.10.021 (2015).

20. Amano, T., Lamming, J. D. & Sutherland, W. J. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience* **66,** 393–400. https://doi.org/10.1093/biosci/biw022 (2016).

21. Sullivan, B. L. *et al.* Using open access observational data for conservation action: A case study for birds. *Biological Conservation* **208,** 5–14. https://doi.org/10.1016/j.biocon.2016.04.031 (2017).

22. Speed, J. D. *et al.* Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PloS one* **13,** e0196417. https://doi.org/10.1371/journal.pone.0196417 (2018).

23. Nelson, G. & Ellis, S. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B* **374,** 20170391. https://doi.org/10.1098/rstb.2017.0391 (2019).

24. Hedrick, B. P. *et al.* Digitization and the future of natural history collections. *BioScience* **70,** 243–251. https://doi.org/10.1093/biosci/biz163 (2020).

25. Miller, S. E. *et al.* Building natural history collections for the twenty-first century and beyond. *BioScience* **70,** 674–687. https://doi.org/10.1093/biosci/biaa069 (2020).

26. Tulloch, A. I., Possingham, H. P., Joseph, L. N., Szabo, J. & Martin, T. G. Realising the full potential of citizen science monitoring programs. *Biological Conservation* **165,** 128–138. https://doi.org/10.1016/j.biocon.2013.05.025 (2013).

27. Bayraktarov, E. *et al.* Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution,* 239. https://doi.org/10.3389/fevo.2018.00239 (2019).

28. James, S. A. *et al.* Herbarium data: Global biodiversity and societal botanical needs for novel research. *Applications in plant sciences* **6,** e1024. https://doi.org/10.1002/aps3.1024 (2018).

29. Støa, B., Halvorsen, R., Mazzoni, S. & Gusarov, V. Sampling bias in presence-only data used for species distribution modelling: theory and methods for detecting sample bias and its effects on models. *Sommerfeltia* **38,** 1–53. https://doi.org/10.2478/som-2018-0001 (2018).

30. Gelfand, A. E. & Shirota, S. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs* **89,** e01372. https://doi.org/10.1002/ecm.1372 (2019).

31. Grimmett, L., Whitsed, R. & Horta, A. Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecological Modelling* **431,** 109194. https://doi.org/10.1016/j.ecolmodel.2020.109194 (2020).

32. Sicacha-Parada, J., Steinsland, I., Cretois, B. & Borgelt, J. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. *Spatial Statistics* **42,** 100446. https://doi.org/10.1016/j.spasta.2020.100446 (2021).

33. Johnston, A. *et al.* Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions* **27,** 1265–1277. https://doi.org/10.1111/ddi.13271 (2021).

34. Petersen, T. K., Speed, J. D. M., Grøtan, V. & Austrheim, G. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. *Ecological Solutions and Evidence* **2,** e12048. https://doi.org/10.1002/2688-8319.12048 (2021).

35. Tessarolo, G., Ladle, R., Rangel, T. & Hortal, J. Temporal degradation of data limits biodiversity research. *Ecology and Evolution* **7,** 6863–6870. https://doi.org/10.1002/ece3.3259 (2017).

36. Mesibov, R. An audit of some processing effects in aggregated occurrence records. *ZooKeys,* 129. https://doi.org/10.3897/zookeys.751.24791 (2018).

37. Escribano, N., Ariño, A. H. & Galicia, D. Biodiversity data obsolescence and land uses changes. *PeerJ* **4,** e2743. https://doi.org/10.7717/peerj.2743 (2016).

38. Araújo, M. B. *et al.* Standards for distribution models in biodiversity assessments. *Science Advances* **5,** eaat4858. https://doi.org/10.1126/sciadv.aat4858 (2019).

39. Kelling, S. *et al.* Using Semistructured Surveys to Improve Citizen Science Data for Monitoring Biodiversity. *BioScience* **69,** 170–179. https://doi.org/10.1093/biosci/biz010 (Mar. 2019).

40. Huettmann, F. in *Data Mining for Global Trends in Mountain Biodiversity* 25–28 (CRC Press, 2009). https://doi.org/10.1201/9781420083705.ch4.

41. Dobson, A. D. *et al.* Making messy data work for conservation. *One Earth* **2,** 455–465. https://doi.org/10.1016/j.oneear.2020.04.012 (2020).

42. Foster, S. D. *et al.* Effects of ignoring survey design information for data reuse. *Ecological Applications* **31,** e02360. https://doi.org/10.1002/eap.2360 (2021).

43. Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D. & Peres-Neto, P. Ecological data should not be so hard to find and reuse. *Trends in ecology & evolution* **34,** 494–496. https://doi.org/10.1016/j.tree.2019.04.005 (2019).

44. Bishop, B. W., Hank, C., Webster, J. & Howard, R. Scientists' data discovery and reuse behavior:(Meta) data fitness for use and the FAIR data principles. *Proceedings of the Association for Information Science and Technology* **56,** 21–31. https://doi.org/10.1002/pra2.4 (2019).

45. Escribano, N., Galicia, D. & Ariño, A. H. The tragedy of the biodiversity data commons: a data impediment creeping nigher? *Database* **2018.** https://doi.org/10.1093/database/bay033 (2018).

46. Luo, M. *et al.* The use of Global Biodiversity Information Facility (GBIF)-mediated data in publications written in Chinese. *Global Ecology and Conservation* **25,** e01406. https://doi.org/10.1016/j.gecco.2020.e01406 (2021).

47. Ariño, A. H. Approaches to estimating the universe of natural history collections data. *Biodiversity informatics* **7.** https://doi.org/10.17161/bi.v7i2.3991 (2010).

48. Aria, M. & Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics* **11,** 959–975. https://doi.org/10.1016/j.joi.2017.08.007 (2017).

49. Westgate, M. J. revtools: An R package to support article screening for evidence synthesis. *Research synthesis methods* **10,** 606–614. https://doi.org/10.1002/jrsm.1374 (2019).

50. Asmussen, C. B. & Møller, C. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data* **6,** 1–18. https://doi.org/10.1186/s40537-019-0255-7 (2019).

51. Hao, T., Elith, J., Guillera-Arroita, G. & Lahoz-Monfort, J. J. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Diversity and Distributions* **25,** 839–852. https://doi.org/10.1111/ddi.12892 (2019).

52. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2020). https://www.R-project.org/.

53. Mandeville, C. Open data practices among users of primary biodiversity data. https://doi.org/10.17605/OSF.IO/JUEQC (2021).

54. Stork, H., Astrin, J. J., *et al.* Trends in biodiversity research—a bibliometric assessment. *Open Journal of Ecology* **4,** 354. http://doi.org/10.4236/oje.2014.47033 (2014).

55. Li, K., Greenberg, J. & Dunic, J. Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. *Journal of the Association for Information Science and Technology* **71,** 172–182. https://doi.org/10.1002/asi.24226 (2020).

56. Guisan, A. *et al.* Predicting species distributions for conservation decisions. *Ecology letters* **16,** 1424–1435. https://doi.org/10.1111/ele.12189 (2013).

57. Tulloch, A. I. *et al.* A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nature ecology & evolution* **2,** 1209–1217. https://doi.org/10.1038/s41559-018-0608-1 (2018).

58. Fanelli, D. & Larivière, V. Researchers' individual publication rate has not increased in a century. *PloS one* **11,** e0149504. https://doi.org/10.1371/journal.pone.0149504 (2016).

59. GBIF. GBIF science review 2019. *Global Biodiversity Information Facility.* https://doi.org/10.15468/QXXG-7K93 (2019).

60. Tydecks, L., Jeschke, J. M., Wolf, M., Singer, G. & Tockner, K. Spatial and topical imbalances in biodiversity research. *PloS one* **13,** e0199327. https://doi.org/10.1371/journal.pone.0199327 (2018).

61. Serra-Diaz, J. M., Enquist, B. J., Maitner, B., Merow, C. & Svenning, J.-C. Big data of tree species distributions: how big and how good? *Forest Ecosystems* **4,** 1–12. https://doi.org/10.1186/s40663-017-0120-0 (2017).

62. Pelayo-Villamil, P. *et al.* Completeness of national freshwater fish species inventories around the world. *Biodiversity and Conservation* **27,** 3807–3817. https://doi.org/10.1007/s10531-018-1630-y (2018).

63. Nuñez, M. A. *et al.* Assessing the uneven global distribution of readership, submissions and publications in applied ecology: Obvious problems without obvious solutions. https://doi.org/10.1111/1365-2664.13319 (2019).

64. Ramirez, K. S. *et al.* The future of ecology is collaborative, inclusive and deconstructs biases. *Nature Ecology & Evolution* **2,** 200–200. https://doi.org/10.1038/s41559-017-0445-7 (2018).

65. Pettorelli, N. *et al.* How international journals can support ecology from the Global South. *Journal of Applied Ecology* **58,** 4–8. https://doi.org/10.1111/1365-2664.13815 (2021).

66. Boshoff, N. Neo-colonialism and research collaboration in Central Africa. *Scientometrics* **81,** 413–434. https://doi.org/10.1007/s11192-008-2211-8 (2009).

67. Habel, J. C. *et al.* Towards more equal footing in north–south biodiversity research: European and sub-Saharan viewpoints. *Biodiversity and conservation* **23,** 3143–3148. https://doi.org/10.1007/s10531-014-0761-z (2014).

68. Di Marco, M. *et al.* Changing trends and persisting biases in three decades of conservation science. *Global Ecology and Conservation* **10,** 32–42. https://doi.org/10.1016/j.gecco.2017.01.008 (2017).

69. Serwadda, D., Ndebele, P., Grabowski, M. K., Bajunirwe, F. & Wanyenze, R. K. Open data sharing and the Global South—Who benefits? *Science* **359,** 642–643. https://doi.org/10.1126/science.aap8395 (2018).

70. Eichhorn, M. P., Baker, K. & Griffiths, M. Steps towards decolonising biogeography. *Frontiers of Biogeography* **12,** 1–7. https://doi.org/10.21425/F5FBG44795 (2020).

71. Trisos, C. H., Auerbach, J. & Katti, M. Decoloniality and anti-oppressive practices for a more ethical ecology. *Nature Ecology & Evolution* **5,** 1205–1212. https://doi.org/10.1038/s41559-021-01460-w (May 2021).

72. Dias, D. *et al.* Repatriation Data: More than two million species occurrence records added to the Brazilian Biodiversity Information Facility Repository (SiBBr). *Biodiversity Data Journal.* https://doi.org/10.3897/BDJ.5.e12012 (2017).

73. Parsons, A. W., Goforth, C., Costello, R. & Kays, R. The value of citizen science for ecological monitoring of mammals. *PeerJ* **6,** e4536. https://doi.org/10.7717/peerj.4536 (2018).

74. Pino-Del-Carpio, A., Ariño, A. H., Villarroya, A., Puig, J. & Miranda, R. The biodiversity data knowledge gap: Assessing information loss in the management of Biosphere Reserves. *Biological Conservation* **173,** 74–79. https://doi.org/10.1016/j.biocon.2013.11.020 (2014).

75. Chambers, L. E. *et al.* Southern hemisphere biodiversity and global change: data gaps and strategies. *Austral Ecology* **42,** 20–30. https://doi.org/10.1111/aec.12391 (2017).

76. Ondei, S., Brook, B. W. & Buettel, J. C. Nature's untold stories: an overview on the availability and type of on-line data on long-term biodiversity monitoring. *Biodiversity and Conservation* **27,** 2971–2987. https://doi.org/10.1007/s10531-018-1582-2 (2018).

77. Wetzel, F. T. *et al.* Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in Europe. *Biological conservation* **221,** 78–85. https://doi.org/10.1016/j.biocon.2017.12.024 (2018).

78. Hochkirch, A. *et al.* A strategy for the next decade to address data deficiency in neglected biodiversity. *Conservation Biology* **35,** 502–509. https://doi.org/10.1111/cobi.13589 (2021).

79. Stephenson, P. *et al.* Unblocking the flow of biodiversity data for decision-making in Africa. *Biological Conservation* **213,** 335–340. https://doi.org/10.1016/j.biocon.2016.09.003 (2017).

80. Cretois, B., Linnell, J. D., Grainger, M., Nilsen, E. B. & Rød, J. K. Hunters as citizen scientists: Contributions to biodiversity monitoring in Europe. *Global Ecology and Conservation* **23,** e01077. https://doi.org/10.1016/j.gecco.2020.e01077 (2020).

81. Vorisek, P. *et al.* Wetzel et al. fail to identify the real gaps in European bird monitoring. *Biological Conservation* **225,** 245–246. https://doi.org/10.1016/j.biocon.2018.07.001 (2018).

82. Grainger, M. J., Bolam, F. C., Stewart, G. B. & Nilsen, E. B. Evidence synthesis for tackling research waste. *Nature Ecology & Evolution* **4,** 495–497. https://doi.org/10.1038/s41559-020-1141-6 (2020).

83. Faith, D. *et al.* Bridging the biodiversity data gaps: Recommendations to meet users' data needs. *Biodiversity Informatics* **8.** https://doi.org/10.17161/bi.v8i2.4126 (2013).

84. Singer, R. A., Ellis, S. & Page, L. M. Awareness and use of biodiversity collections by fish biologists. *Journal of Fish Biology* **96,** 297–306. https://doi.org/10.1111/jfb.14167 (2020).

85. Curty, R. G., Crowston, K., Specht, A., Grant, B. W. & Dalton, E. D. Attitudes and norms affecting scientists' data reuse. *PloS one* **12,** e0189288. https://doi.org/10.1371/journal.pone.0189288 (2017).

86. Johnson, N., Druckenmiller, M. L., Danielsen, F. & Pulsifer, P. L. The use of digital platforms for community-based monitoring. *BioScience* **71,** 452–466. https://doi.org/10.1093/biosci/biaa162 (2021).

87. Costello, M. J. & Wieczorek, J. Best practice for biodiversity data management and publication. *Biological Conservation* **173,** 68–73. https://doi.org/10.1016/j.biocon.2013.10.018 (2014).

88. Franz, N. M. & Sterner, B. W. To increase trust, change the social design behind aggregated biodiversity data. *Database* **2018.** https://doi.org/10.1093/database/bax100 (2018).

89. Blair, J. *et al.* Towards a catalogue of biodiversity databases: An ontological case study. *Biodiversity Data Journal* **8.** https://doi.org/10.3897/BDJ.8.e32765 (2020).

90. Monfils, A. K. *et al.* Regional collections are an essential component of biodiversity research infrastructure. *BioScience* **70,** 1045–1047. https://doi.org/10.1093/biosci/biaa102 (2020).

91. Vines, T. H. *et al.* The availability of research data declines rapidly with article age. *Current biology* **24,** 94–97. https://doi.org/10.1016/j.cub.2013.11.014 (2014).

92. Calver, M. C., Goldman, B., Hutchings, P. A. & Kingsford, R. T. Why discrepancies in searching the conservation biology literature matter. *Biological Conservation* **213,** 19–26. https://doi.org/10.1016/j.biocon.2017.06.028 (2017).

93. Altman, M. & Crosas, M. The Evolution of Data Citation: From Principles to Implementation. *IASSIST QUARTERLy,* 63. https://doi.org/10.29173/iq504 (2013).

94. Penev, L. *et al.* Strategies and guidelines for scholarly publishing of biodiversity data. *Research Ideas and Outcomes* **3,** e12431. https://doi.org/10.3897/rio.3.e12431 (2017).

95. Groom, Q. *et al.* People are essential to linking biodiversity data. *Database* **2020.** https://doi.org/10.1093/database/baaa072 (2020).

96. Cooper, C. B., Shirk, J. & Zuckerberg, B. The invisible prevalence of citizen science in global research: migratory birds and climate change. *PloS one* **9,** e106508. https://doi.org/10.1371/journal.pone.0106508 (2014).

97. Chandler, M. *et al.* Contribution of citizen science towards international biodiversity monitoring. *Biological conservation* **213,** 280–294. https://doi.org/10.1016/j.biocon.2016.09.004 (2017).

98. Pearce-Higgins, J. W. *et al.* Overcoming the challenges of public data archiving for citizen science biodiversity recording and monitoring schemes. *Journal of Applied Ecology* **55,** 2544–2551. https://doi.org/10.1111/1365-2664.13180 (2018).

99. MacPhail, V. J. & Colla, S. R. Power of the people: A review of citizen science programs for conservation. *Biological Conservation* **249,** 108739. https://doi.org/10.1016/j.biocon.2020.108739 (2020).

100. Mandeville, C. P. & Finstad, A. G. Community science supports research on protected area resilience. *Conservation Science and Practice* **3.** https://doi.org/10.1111/csp2.442 (2021).

101. Vaz, U., Cunha, H. & Nabout, J. Trends and biases in global scientific literature about ecological niche models. *Brazilian Journal of Biology* **75,** 17–24. https://doi.org/10.1590/1519-6984.22713 (2015).

102. Norberg, A. *et al.* A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological monographs* **89,** e01370. https://doi.org/10.1002/ecm.1370 (2019).

103. Zurell, D. *et al.* A standard protocol for reporting species distribution models. *Ecography* **43,** 1261–1277. https://doi.org/10.1111/ecog.04960 (2020).

104. Pacifici, K. *et al.* Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology* **98,** 840–850. https://doi.org/10.1002/ecy.1710 (2017).

105. Fletcher Jr, R. J. *et al.* A practical guide for combining data to model species distributions. *Ecology* **100,** e02710. https://doi.org/10.1002/ecy.2710 (2019).

106. Miller, D. A., Pacifici, K., Sanderlin, J. S. & Reich, B. J. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution* **10,** 22–37. https://doi.org/10.1111/2041-210X.13110 (2019).

107. Isaac, N. J. *et al.* Data integration for large-scale models of species distributions. *Trends in ecology & evolution* **35,** 56–67. https://doi.org/10.1016/j.tree.2019.08.006 (2020).

108. Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. & O'Hara, R. B. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography* **43,** 1413–1422. https://doi.org/10.1111/ecog.05146 (2020).

109. Zipkin, E. F. *et al.* Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment* **19,** 30–38. https://doi.org/10.1002/fee.2290 (2021).

110. Isaac, N. J., van Strien, A. J., August, T. A., de Zeeuw, M. P. & Roy, D. B. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* **5,** 1052–1060. https://doi.org/10.1111/2041-210X.12254 (2014).

111. Joppa, L. N. *et al.* Troubling trends in scientific software use. *Science* **340,** 814–815. https://doi.org/10.1126/science.1231535 (2013).

112. Merow, C., Smith, M. J. & Silander Jr, J. A. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* **36,** 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x (2013).

113. Kass, J. M. *et al.* Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolution* **9,** 1151–1156. https://doi.org/10.1111/2041-210X.12945 (2018).

114. Rapacciuolo, G. Strengthening the contribution of macroecological models to conservation practice. *Global Ecology and Biogeography* **28,** 54–60. https://doi.org/10.1111/geb.12848 (2019).

115. Sofaer, H. R. *et al.* Development and delivery of species distribution models to inform decision-making. *BioScience* **69,** 544–557. https://doi.org/10.1093/biosci/biz045 (2019).

116. Soranno, P. A. *et al.* Ecological prediction at macroscales using big data: Does sampling design matter? *Ecological Applications* **30,** e02123. https://doi.org/10.1002/eap.2123 (2020).

117. Muscatello, A., Elith, J. & Kujala, H. How decisions about fitting species distribution models affect conservation outcomes. *Conservation Biology* **35,** 1309–1320. https://doi.org/10.1111/cobi.13669 (2021).

118. Sillero, N. & Barbosa, A. M. Common mistakes in ecological niche models. *International Journal of Geographical Information Science* **35,** 213–226. https://doi.org/10.1080/13658816.2020.1798968 (2021).

119. Kelling, S. *et al.* Using Semistructured Surveys to Improve Citizen Science Data for Monitoring Biodiversity. *BioScience* **69,** 170–179. https://doi.org/10.1093/biosci/biz010 (Mar. 2019).

120. Roche, D. G., Kruuk, L. E., Lanfear, R. & Binning, S. A. Public data archiving in ecology and evolution: how well are we doing? *PLoS biology* **13,** e1002295. https://doi.org/10.1371/journal.pbio.1002295 (2015).

121. Specht, A., Bolton, M. P., Kingsford, B., Specht, R. L. & Belbin, L. A story of data won, data lost and data re-found: the realities of ecological data preservation. *Biodiversity Data Journal.* https://doi.org/10.3897/BDJ.6.e28073 (2018).

122. Yackulic, C. B. *et al.* Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution* **4,** 236–243. https://doi.org/10.1111/2041-210x.12004 (2013).

123. Hampton, S. E. *et al.* Big data and the future of ecology. *Frontiers in Ecology and the Environment* **11,** 156–162. https://doi.org/10.1890/120103 (2013).

124. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* **115,** 2584–2589. https://doi.org/10.1073/pnas.1708290115 (2018).

125. Hackett, R. A., Belitz, M. W., Gilbert, E. E. & Monfils, A. K. A data management workflow of biodiversity data from the field to data users. *Applications in Plant Sciences* **7,** e11310. https://doi.org/10.1002/aps3.11310 (2019).

126. Ganzevoort, W., van den Born, R. J., Halffman, W. & Turnhout, S. Sharing biodiversity data: citizen scientists' concerns and motivations. *Biodiversity and Conservation* **26,** 2821–2837. https://doi.org/10.1007/s10531-017-1391-z (2017).

127. Groom, Q., Weatherdon, L. & Geijzendorffer, I. R. Is citizen science an open science in the case of biodiversity observations? *Journal of Applied Ecology* **54,** 612–617. https://doi.org/10.1111/1365-2664.12767 (2017).

128. Fox, R. *et al.* Opinions of citizen scientists on open access to UK butterfly and moth occurrence data. *Biodiversity and Conservation* **28,** 3321–3341. https://doi.org/10.1007/s10531-019-01824-6 (2019).

129. Larson, L. R. *et al.* The diverse motivations of citizen scientists: Does conservation emphasis grow as volunteer participation progresses? *Biological Conservation* **242,** 108428. https://doi.org/10.1016/j.biocon.2020.108428 (2020).

130. Cousijn, H. *et al.* A data citation roadmap for scientific publishers. *Scientific Data* **5.** https://doi.org/10.1038/sdata.2018.259 (2018).

131. Mooney, H. & Newton, M. P. The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication* **1.** https://doi.org/10.7710/2162-3309.1035 (2012).

132. Fenner, M. *et al.* A data citation roadmap for scholarly data repositories. *Scientific Data* **6.** https://doi.org/10.1038/s41597-019-0031-8 (2019).

133. Powers, S. M. & Hampton, S. E. Open science, reproducibility, and transparency in ecology. *Ecological Applications* **29,** e01822. https://doi.org/10.1002/eap.1822 (2019).

134. Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. & Goudie, S. Developing a research data policy framework for all journals and publishers. *Data Science Journal* **19.** https://doi.org/10.5334/dsj-2020-017 (2020).

135. Clements, H. S. *et al.* Fairness and transparency are required for the inclusion of privately protected areas in publicly accessible conservation databases. *Land* **7,** 96. https://doi.org/10.3390/land7030096 (2018).

136. Chapman, A. Current Best Practices for Generalizing Sensitive Species Occurrence Data. https://doi.org/10.15468/DOC-5JP4-5G10 (2020).

137. Carroll, S. R., Herczog, E., Hudson, M., Russell, K. & Stall, S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data* **8,** 1–6. https://doi.org/10.1038/s41597-021-00892-0 (2021).

138. Rubert-Nason, K. *et al.* Ecologist engagement in translational science is imperative for building resilience to global change threats. *Rethinking Ecology* **6,** 65. https://doi.org/10.3897/rethinkingecology.6.64103 (2021).

139. Research Data Alliance International Indigenous Data Sovereignty Interest Group. CARE Principles for Indigenous Data Governance (Sept. 2019).

140. Lynn, S. J., Kaplan, N., Newman, S., Scarpino, R. & Newman, G. Designing a platform for ethical citizen science: A case study of CitSci. org. *Citizen science* **4.** https://doi.org/10.5334/cstp.227 (2019).

141. Tengö, M., Austin, B. J., Danielsen, F. & Fernández-Llamazares, Á. Creating synergies between citizen science and Indigenous and local knowledge. *BioScience* **71,** 503–518. https://doi.org/10.1093/biosci/biab023 (2021).

142. Tenopir, C. *et al.* Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PloS one* **15,** e0229003. https://doi.org/10.1371/journal.pone.0229003 (2020).

143. Soeharjono, S. & Roche, D. G. Reported individual costs and benefits of sharing open data among Canadian Academic Faculty in ecology and evolution. *BioScience* **71,** 750–756. https://doi.org/10.1093/biosci/biab024 (2021).

144. Michener, W. K. Ten simple rules for creating a good data management plan. *PLoS computational biology* **11,** e1004525. https://doi.org/10.1371/journal.pcbi.1004525 (2015).

145. Osawa, T. Perspectives on biodiversity informatics for ecology. *Ecological Research* **34,** 446–456. https://doi.org/10.1111/1440-1703.12023 (2019).

146. Shin, N. *et al.* Toward more data publication of long-term ecological observations. *Ecological Research* **35,** 700–707. https://doi.org/10.1111/1440-1703.12115 (2020).

147. Huang, X. *et al.* Willing or unwilling to share primary biodiversity data: results and implications of an international survey. *Conservation Letters* **5,** 399–406. https://doi.org/10.1111/j.1755-263X.2012.00259.x (2012).

148. Mills, J. A. *et al.* Archiving primary data: solutions for long-term studies. *Trends in Ecology & Evolution* **30,** 581–589. https://doi.org/10.1016/j.tree.2015.07.006 (2015).

149. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. & McGillivray, B. The citation advantage of linking publications to research data. *PloS one* **15,** e0230416. https://doi.org/10.1371/journal.pone.0230416 (2020).

150. Walters, W. H. Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights* **33.** https://doi.org/10.1629/uksg.510 (2020).

151. Chavan, V. & Penev, L. The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Informatics* **12 (Supp 15).** https://doi.org/10.1186/1471-2105-12-S15-S2 (2011).

152. Kattge, J., Díaz, S. & Wirth, C. Of carrots and sticks. *Nature Geoscience* **7,** 778–779. https://doi.org/10.1038/ngeo2280 (2014).

153. Aubin, I. *et al.* Managing data locally to answer questions globally: The role of collaborative science in ecology. *Journal of Vegetation Science* **31,** 509–517. https://doi.org/10.1111/jvs.12864 (2020).

154. Bahlai, C. *et al.* Open science isn't always open to all scientists. *American Scientist* **107,** 78–82. https://doi.org/10.1511/2019.107.2.78 (2019).

155. Nakagawa, S. *et al.* A new ecosystem for evidence synthesis. *Nature Ecology & Evolution* **4,** 498–501. https://doi.org/10.1038/s41559-020-1153-2 (2020).

156. Callaghan, C. T. *et al.* Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience* **71,** 55–63. ISSN: 0006-3568. https://doi.org/10.1093/biosci/biaa131 (Nov. 2020).

157. Ross-Hellauer, T., Deppe, A. & Schmidt, B. Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. *PloS one* **12,** e0189311. https://doi.org/10.1371/journal.pone.0189311 (2017).

158. Hampton, S. E. *et al.* The Tao of open science for ecology. *Ecosphere* **6,** art120. https://doi.org/10.1890/es14-00402.1 (2015).

159. Boland, M. R., Karczewski, K. J. & Tatonetti, N. P. Ten simple rules to enable multi-site collaborations through data sharing. *PLoS Computational Biology* **13,** e1005278. https://doi.org/10.1371/journal.pcbi.1005278 (2017).

160. GoFair Initiative. FAIR principles (2016).

Open data practices among users of primary biodiversity data

SUPPLEMENTARY MATERIALS

S1. Extended methods for literature search and screening.

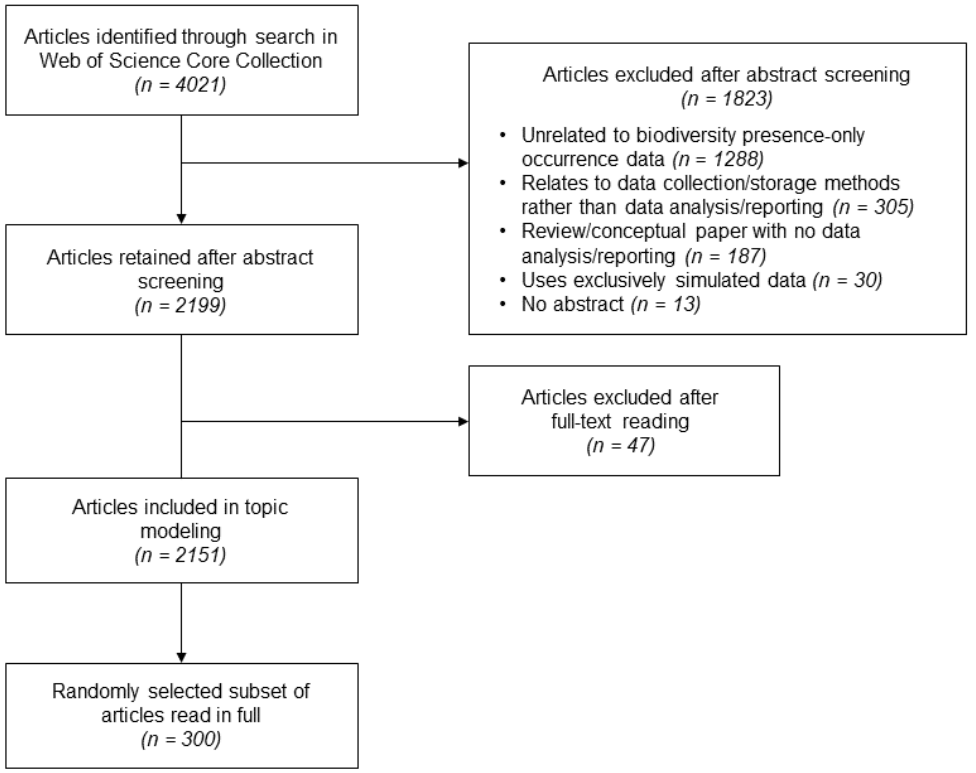
We searched the Web of Science Core Collection to target all scholarly articles that report on the application of presence-only biodiversity occurrence data, targeting articles whose titles, abstracts, or keywords contained any of 31 terms commonly used in the literature to indicate presence-only data as well as any of five terms used to indicate biodiversity:

```
(((TS=("presence-only" OR "presence only" OR "opportunistic
    observation*" OR "opportunistic species observation*" OR
    "opportunistic occurrence*" OR "opportunistic distribution*" OR
    "opportunistic species occurrence*" OR "opportunistic species
    distribution*" OR "pseudo-absence*" OR "pseudoabsence*" OR "inferred
    absence*" OR "presence-background" OR "presence background" OR
    "citizen science" OR "community science" OR "participatory science"
    or "ad hoc data" OR "ad hoc collection" OR "ad hoc method*" OR
    "incidental data" OR "incidental sighting*" OR "incidentally
    collected" OR "geographic one-class data" OR "incidental detection*"
    OR "opportunistic detection*" OR "primary biodiversity data*" OR
    "occurrence record*" OR "atlas data" OR "unstructured occurrence
    data" OR "unstructured species observation" OR "unstructured
    biodiversity data"))
AND (TS=("distribution" OR "species" OR "biodiversity" OR "habitat*" OR
    "niche*")))
AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All
    years
```

The search, conducted on January 4, 2021, returned 4021 peer-reviewed English-language articles.

We screened the abstracts of all returned articles and retained those that demonstrated the analysis or reporting of presence-only occurrence data. in the following categories were excluded: 1) articles unrelated to use of presence-only

biodiversity occurrence data; 2) review or conceptual articles that did not perform data analysis or reporting; 3) articles that focused on the storage or management, rather than analysis or reporting, of occurrence data; and 4) articles that used exclusively simulated data. The article screening process is report in the following diagram, modified from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) scheme[1]. After screening, a total of 2151 articles were included in the review. Data management and bibliometric summary statistics were conducted in part with the bibliometrix package in R[2].

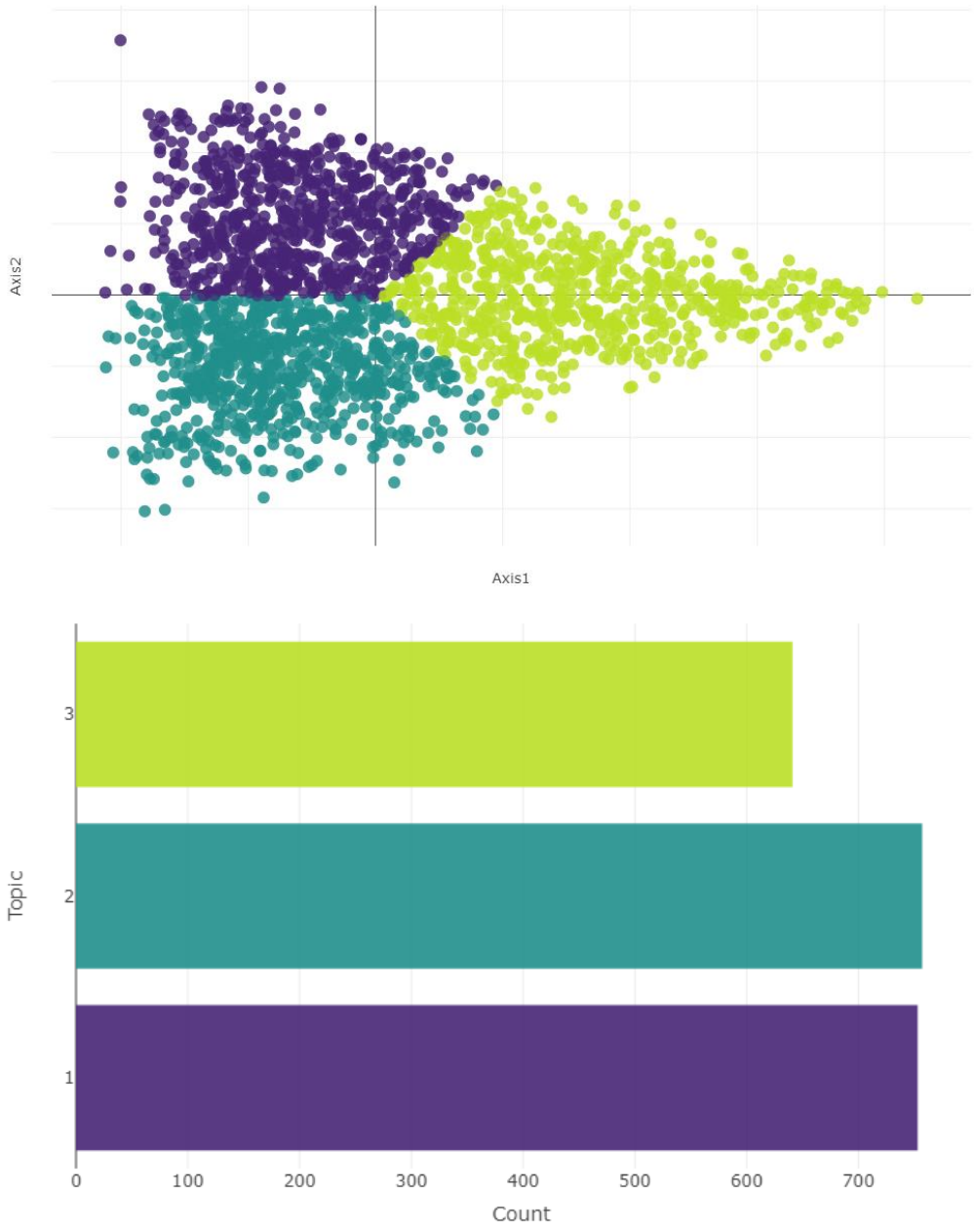


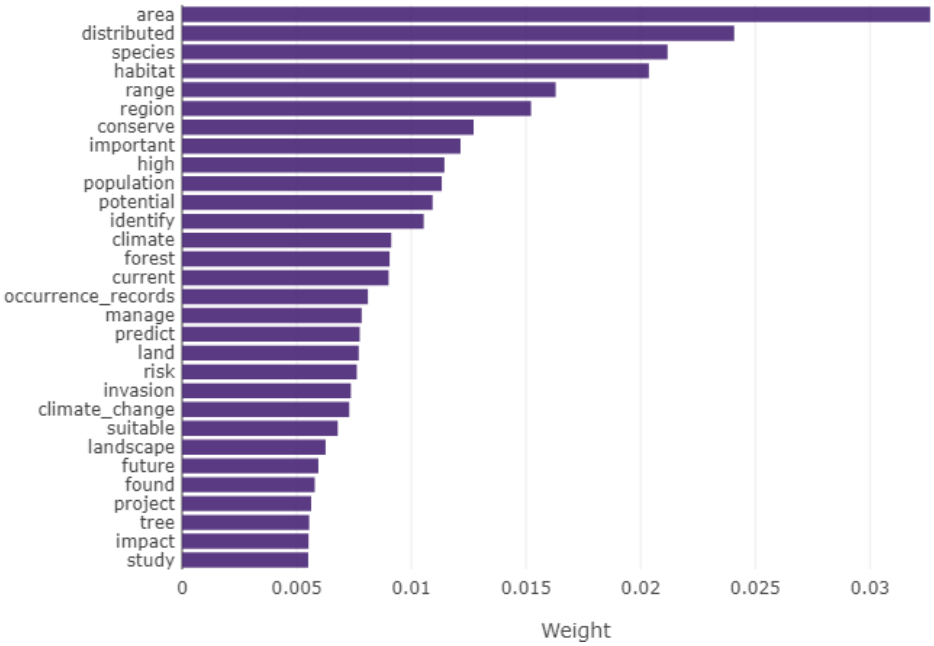
S2. Sets of topic clusters produced by LDA topic modeling.

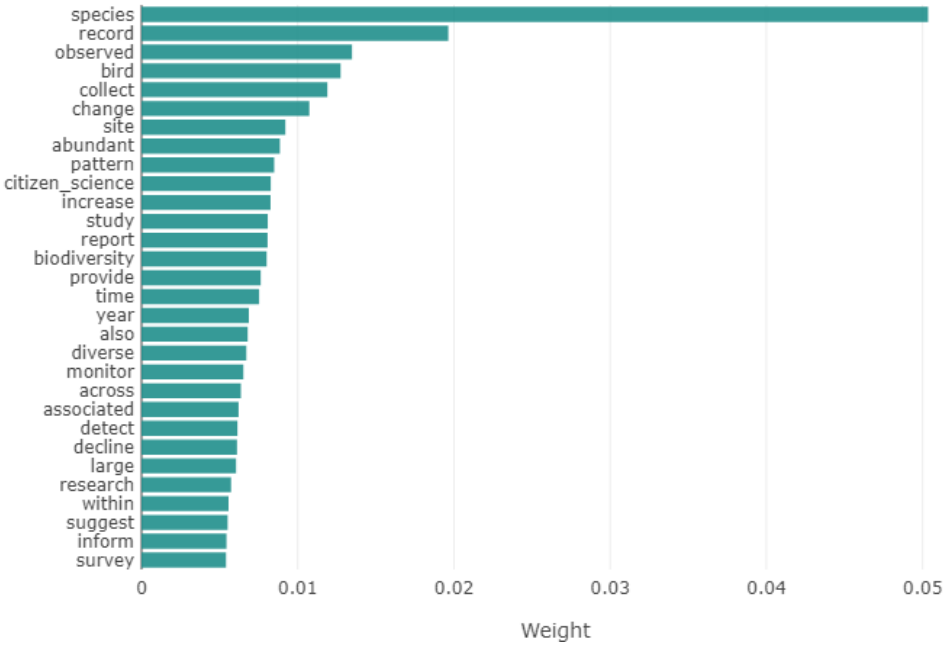
We ran Latent Dirichlet Allocation (LDA) topic modeling six times to produce sets of clusters ranging from three through eight clusters per set. We assessed each set of results for redundancy and interpretability and selected the set of three clusters as the most interpretable and least redundant. All six sets of modeling results are shown here. All topic modeling was conducted and LDA figures were produced using the revtools package in R[3].

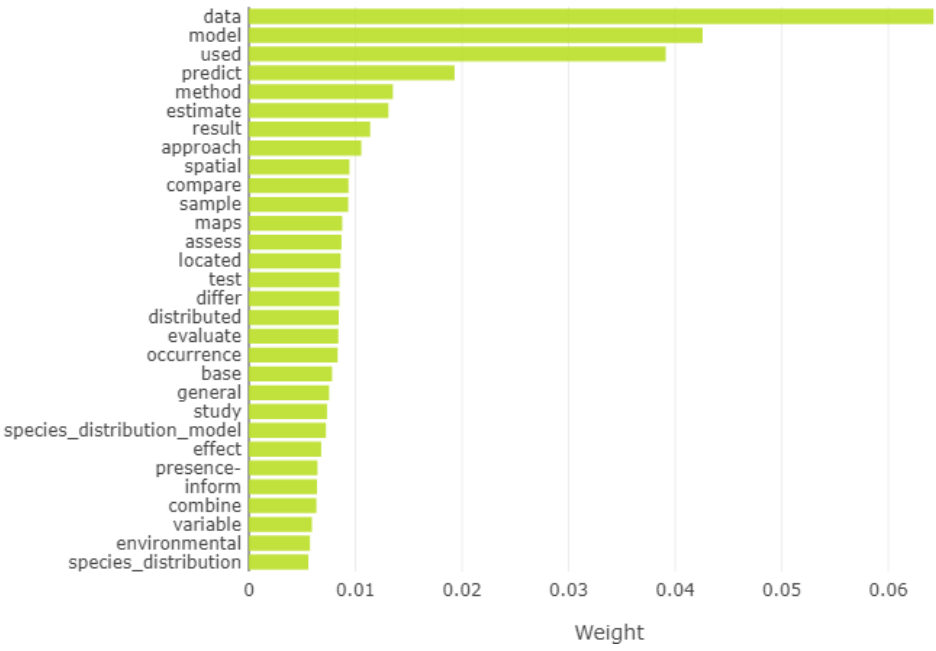
For each set of clusters, the biplot indicates the arrangement of articles relative to each other in terms of topic similarity. Each point represents an article and proximity indicates topical similarity. Colors indicate clusters. The topic bar charts represent the number of articles classified into each topic cluster. The word bar charts indicate the words most strongly associated with each cluster; these word associations were used to assign interpretations to each cluster.

Three clusters

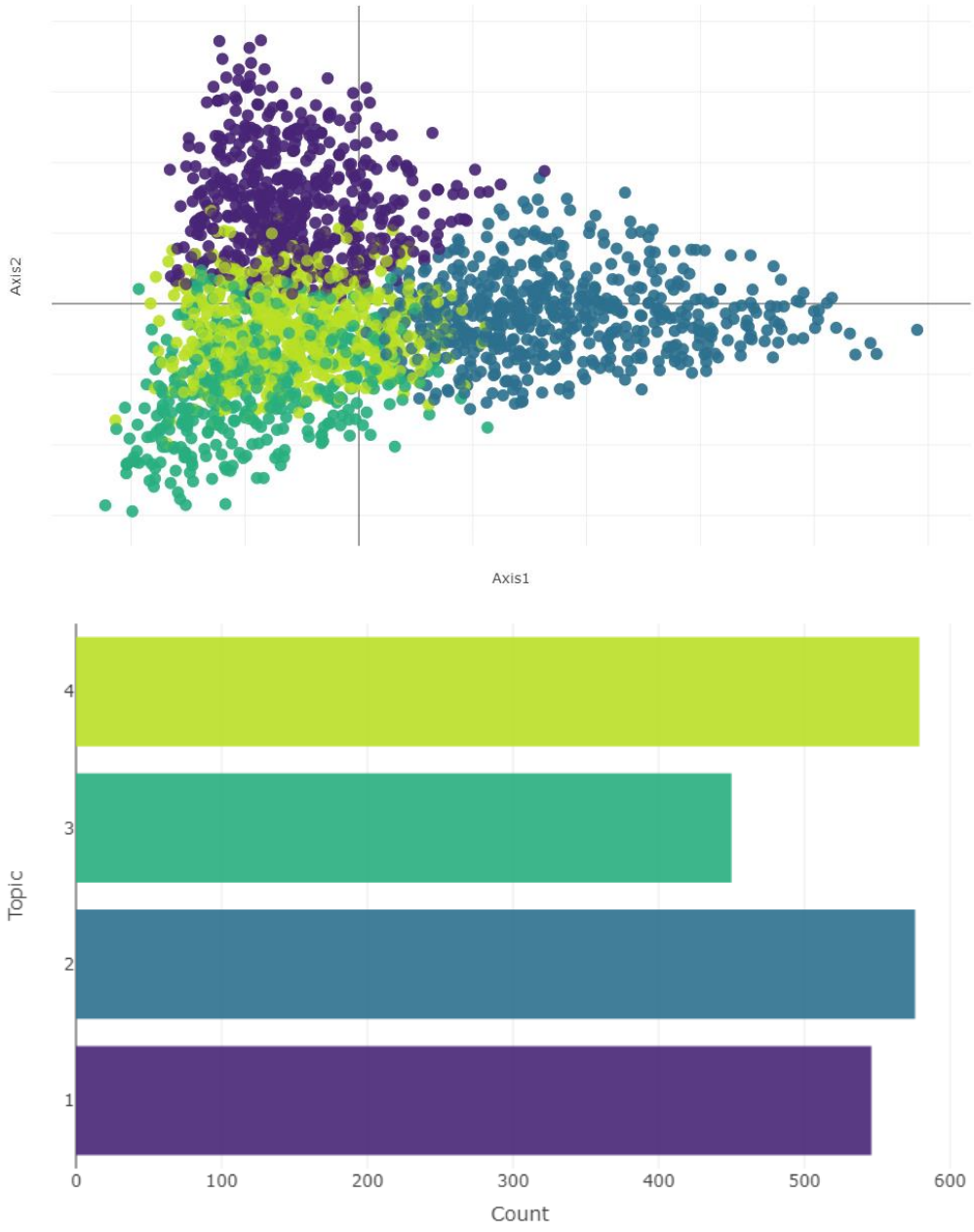


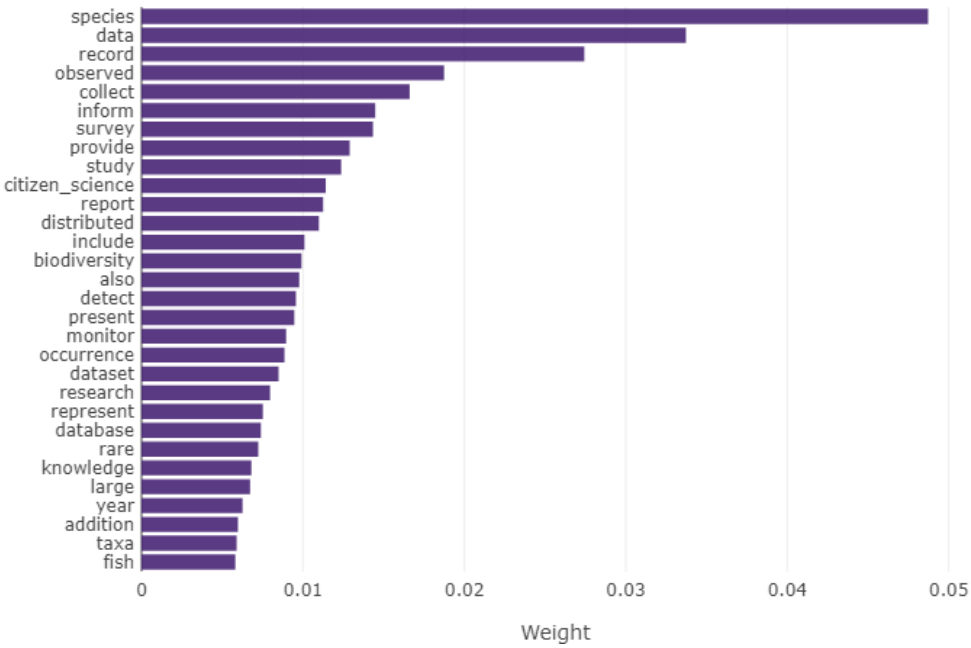

area
distributed
species
habitat
range
region
conserve
important
high
population
potential
identify
climate
forest
current
occurrence_records
manage
predict
land
risk
invasion
climate_change
suitable
landscape
future
found
project
tree
impact
study
0
0.005
0.01
0.015
0.02
0.025
0.03
Weight

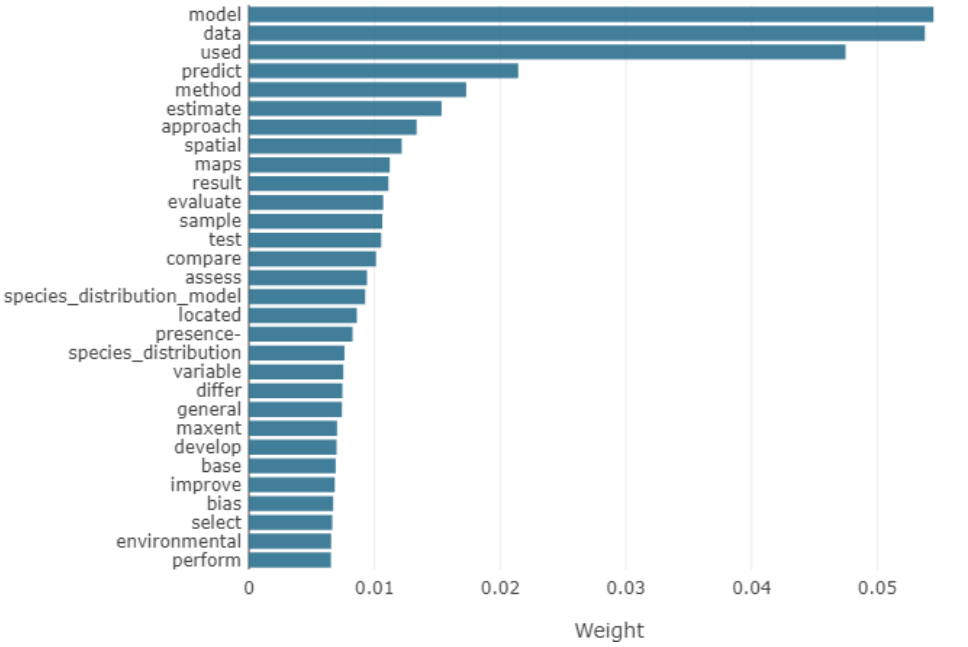

species
record
observed
bird
collect
change
site
abundant
pattern
citizen_science
increase
study
report
biodiversity
provide
time
year
also
diverse
monitor
across
associated
detect
decline
large
research
within
suggest
inform
survey
0
0.01
0.02
0.03
0.04
0.05
Weight

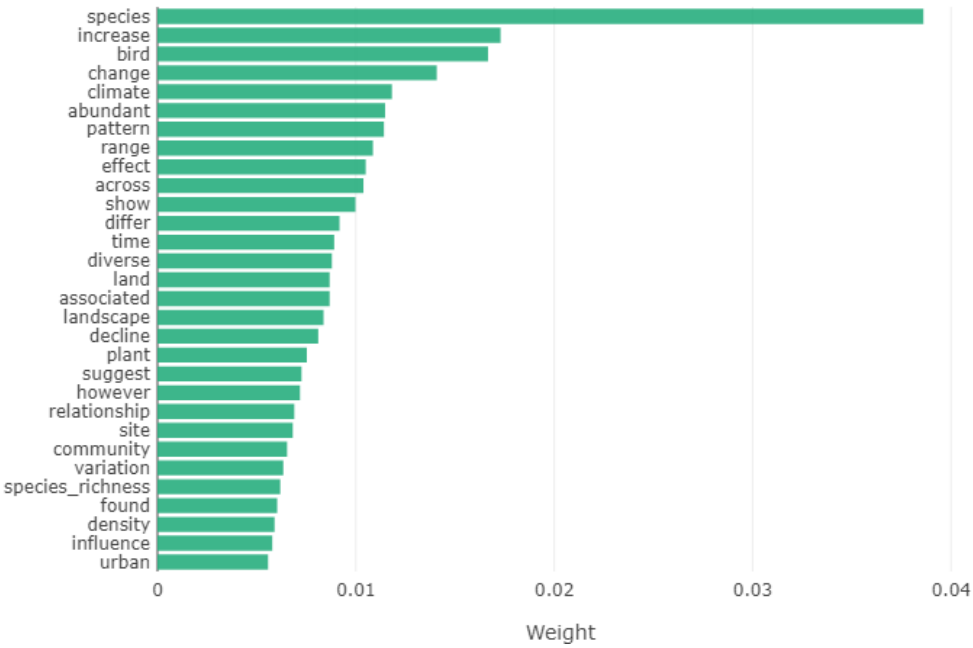

data
model
used
predict
method
estimate
result
approach
spatial
compare
sample
maps
assess
located
test
differ
distributed
evaluate
occurrence
base
general
study
species_distribution_model
effect
presence-
inform
combine
variable
environmental
species_distribution
0
0.01
0.02
0.03
0.04
0.05
0.06
Weight

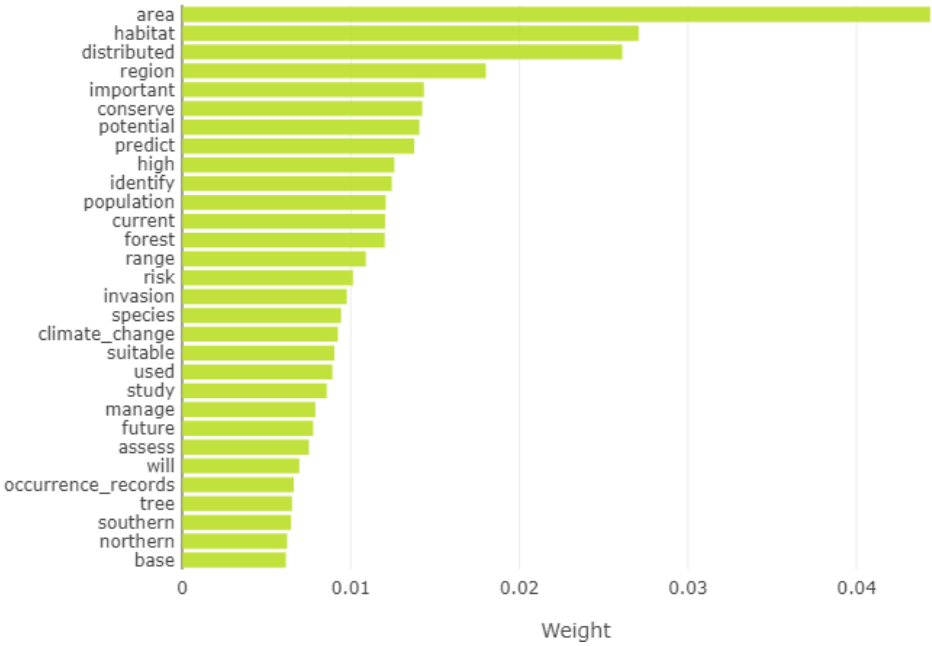
Four clusters



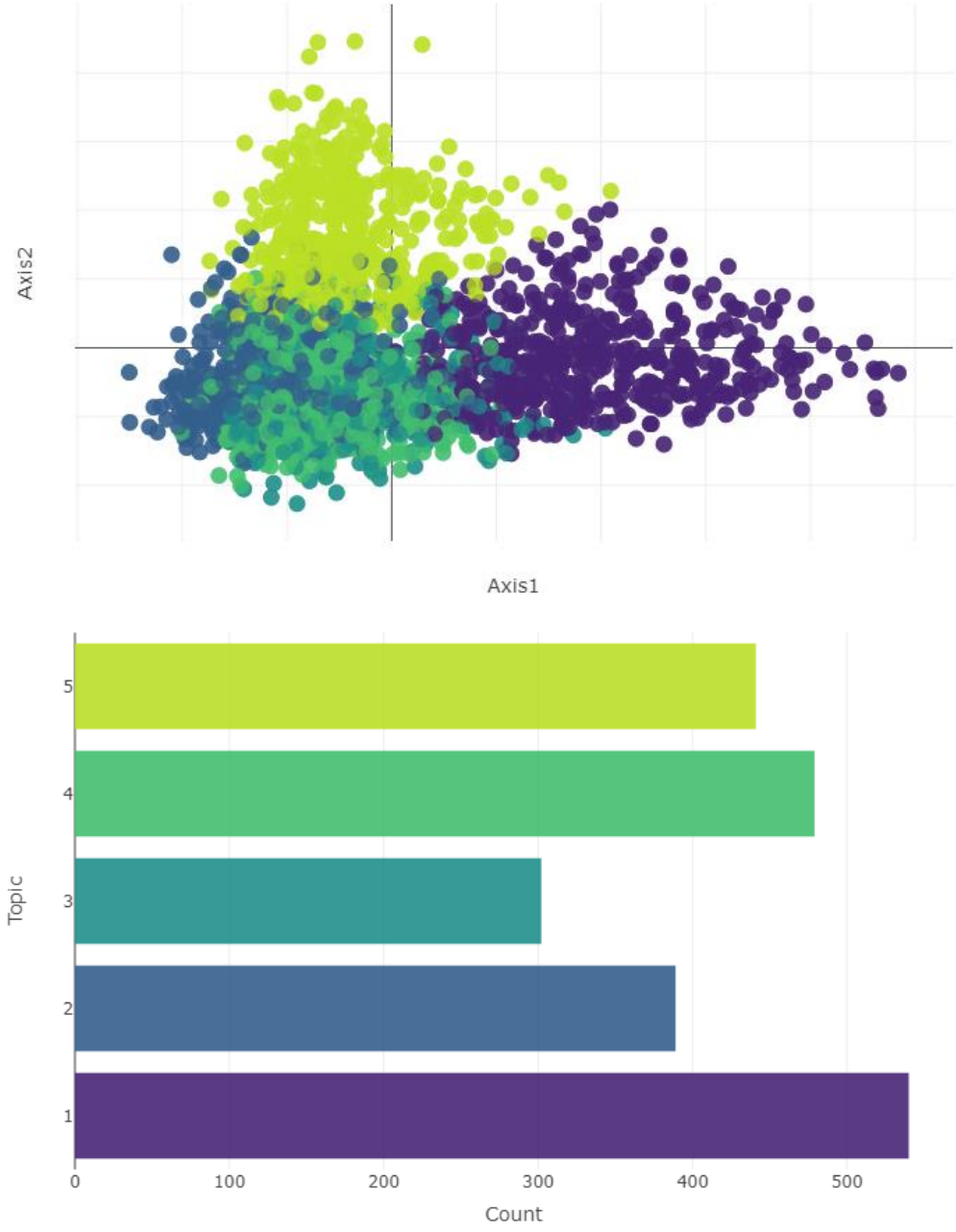

species
data
record
observed
collect
inform
survey
provide
study
citizen_science
report
distributed
include
biodiversity
also
detect
present
monitor
occurrence
dataset
research
represent
database
rare
knowledge
large
year
addition
taxa
fish
0
0.01
0.02
0.03
0.04
0.05
Weight

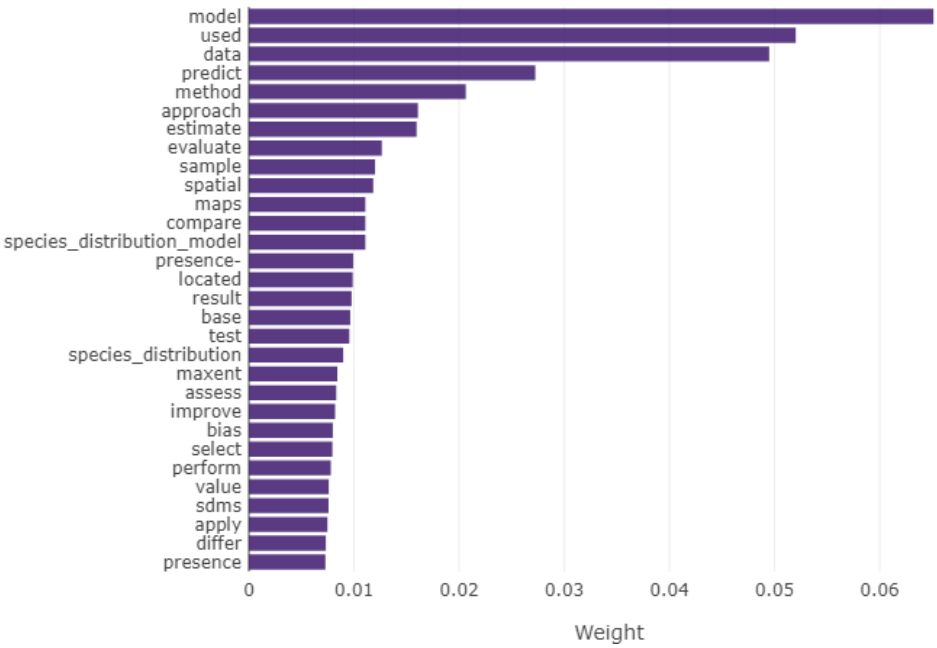

model
data
used
predict
method
estimate
approach
spatial
maps
result
evaluate
sample
test
compare
assess
species_distribution_model
located
presence-
species_distribution
variable
differ
general
maxent
develop
base
improve
bias
select
environmental
perform
0
0.01
0.02
0.03
0.04
0.05
Weight

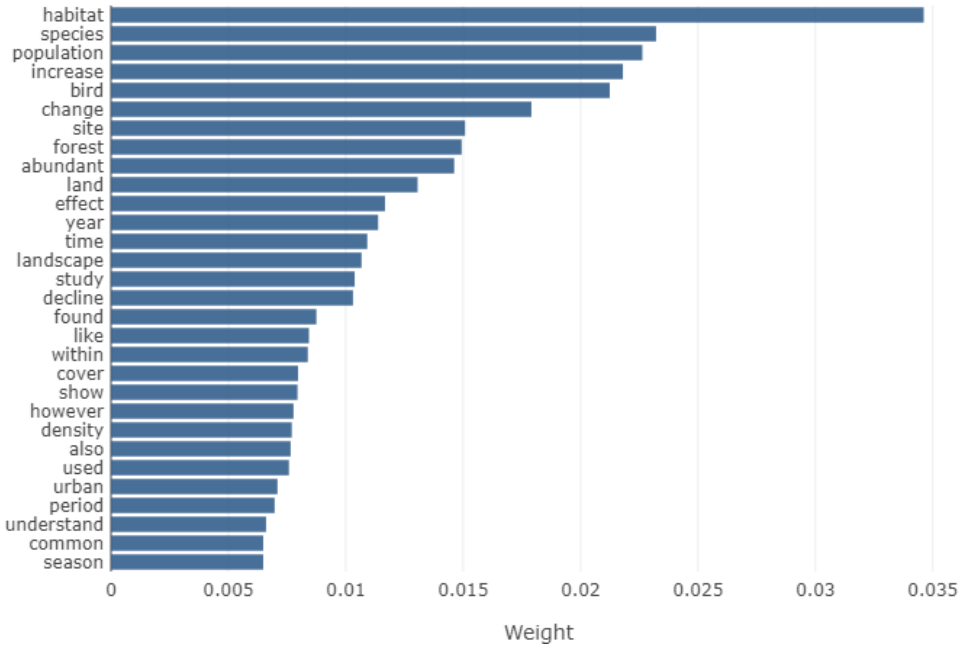

species
increase
bird
change
climate
abundant
pattern
range
effect
across
show
differ
time
diverse
land
associated
landscape
decline
plant
suggest
however
relationship
site
community
variation
species_richness
found
density
influence
urban
0
0.01
0.02
0.03
0.04
Weight

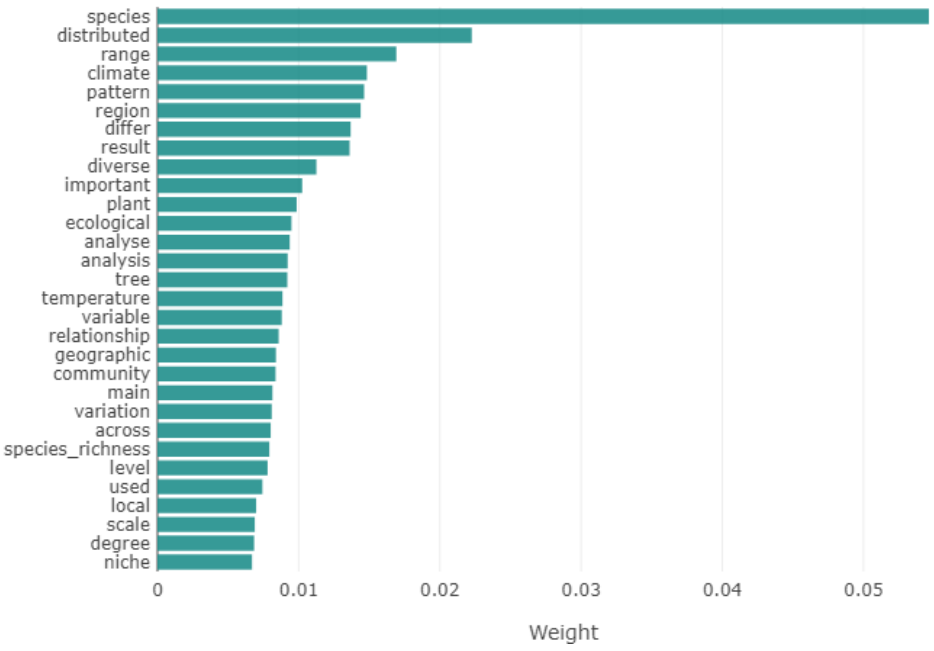

area
habitat
distributed
region
important
conserve
potential
predict
high
identify
population
current
forest
range
risk
invasion
species
climate_change
suitable
used
study
manage
future
assess
will
occurrence_records
tree
southern
northern
base
0
0.01
0.02
0.03
0.04
Weight

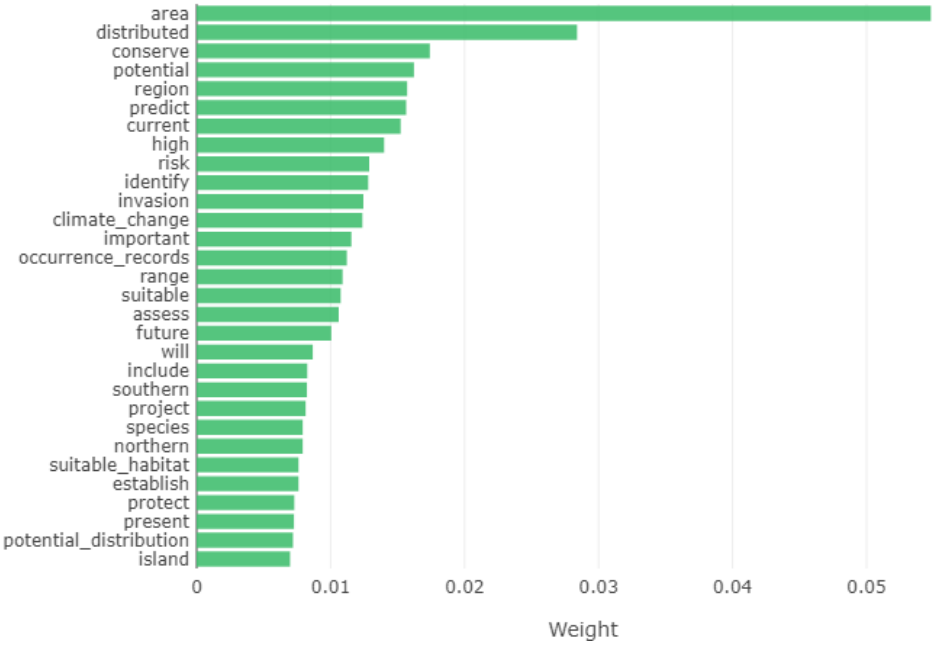
Five clusters

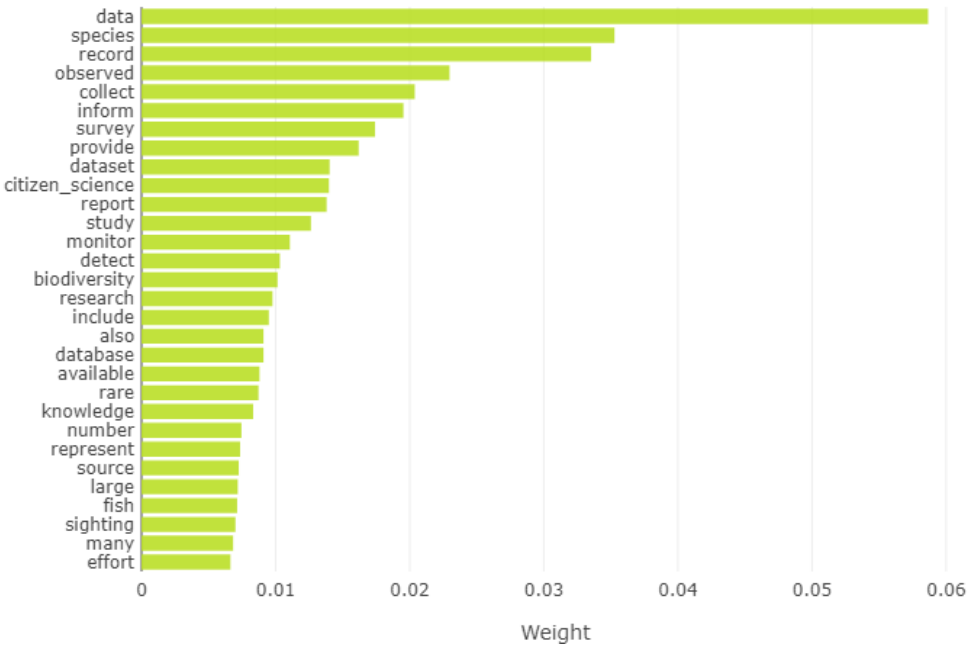



model
used
data
predict
method
approach
estimate
evaluate
sample
spatial
maps
compare
species_distribution_model
presence-
located
result
base
test
species_distribution
maxent
assess
improve
bias
select
perform
value
sdms
apply
differ
presence
0
0.01
0.02
0.03
0.04
0.05
0.06
Weight

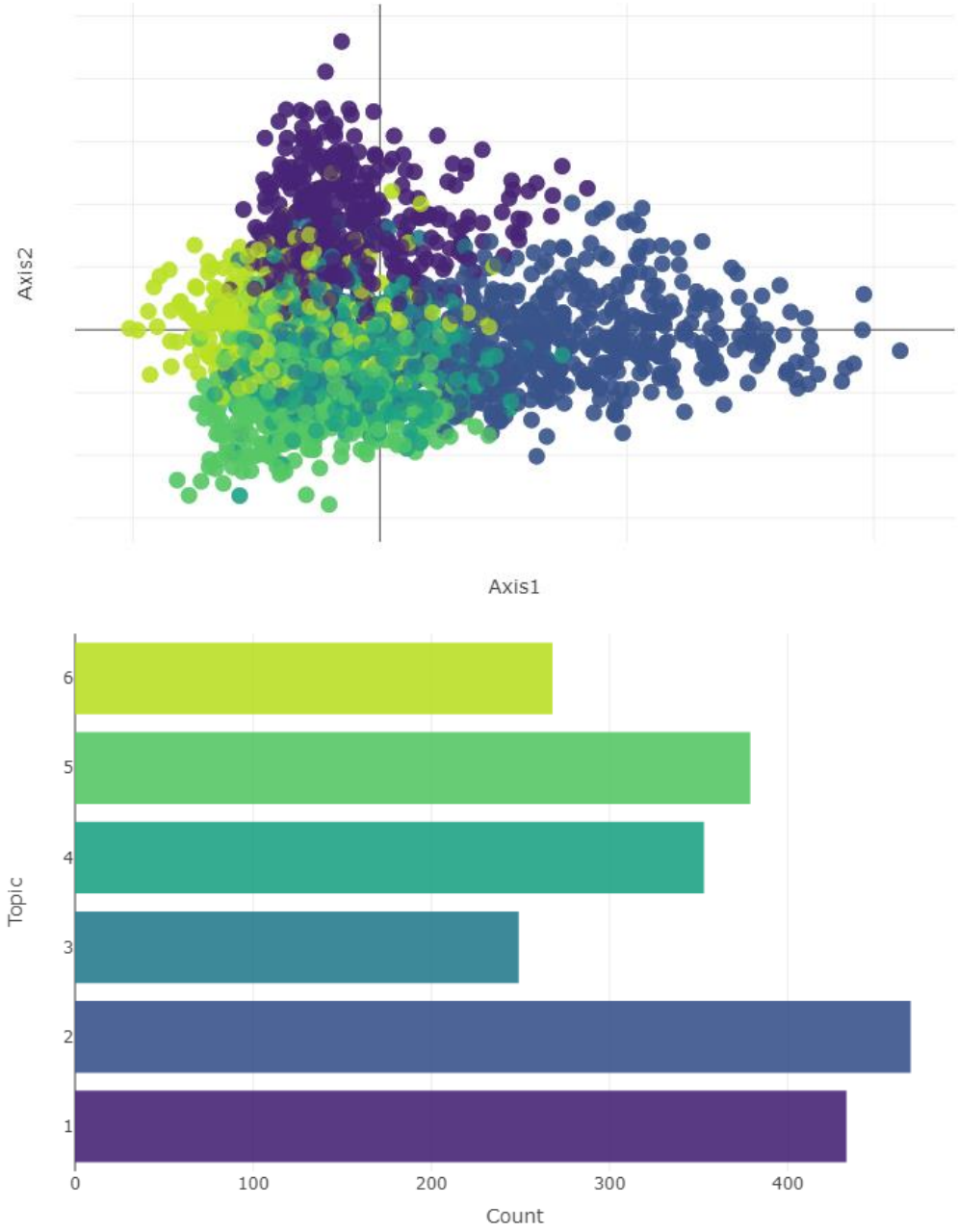

habitat
species
population
increase
bird
change
site
forest
abundant
land
effect
year
time
landscape
study
decline
found
like
within
cover
show
however
density
also
used
urban
period
understand
common
season
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
Weight

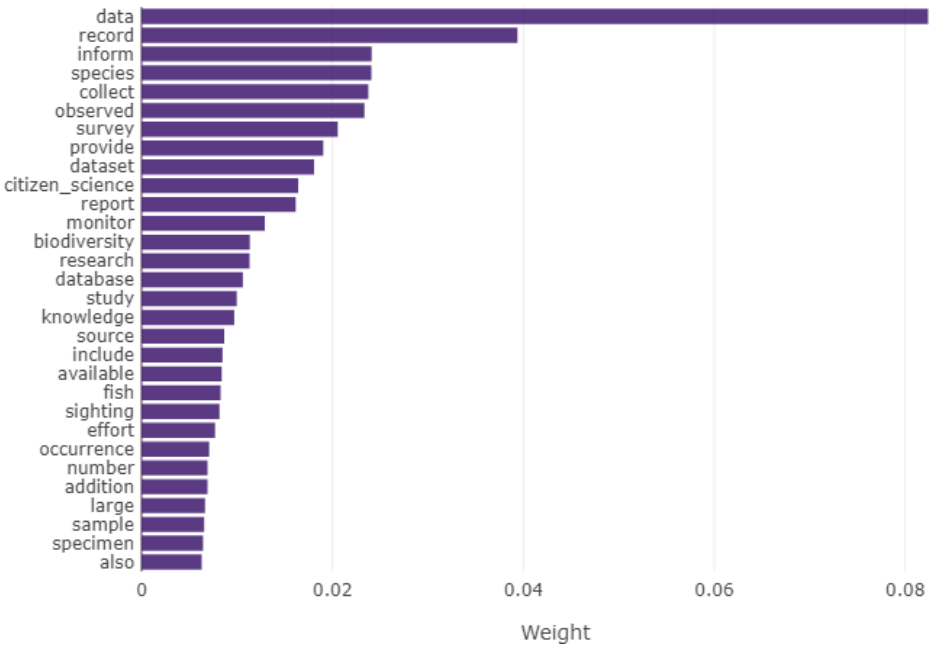

species
distributed
range
climate
pattern
region
differ
result
diverse
important
plant
ecological
analyse
analysis
tree
temperature
variable
relationship
geographic
community
main
variation
across
species_richness
level
used
local
scale
degree
niche
0
0.01
0.02
0.03
0.04
0.05
Weight

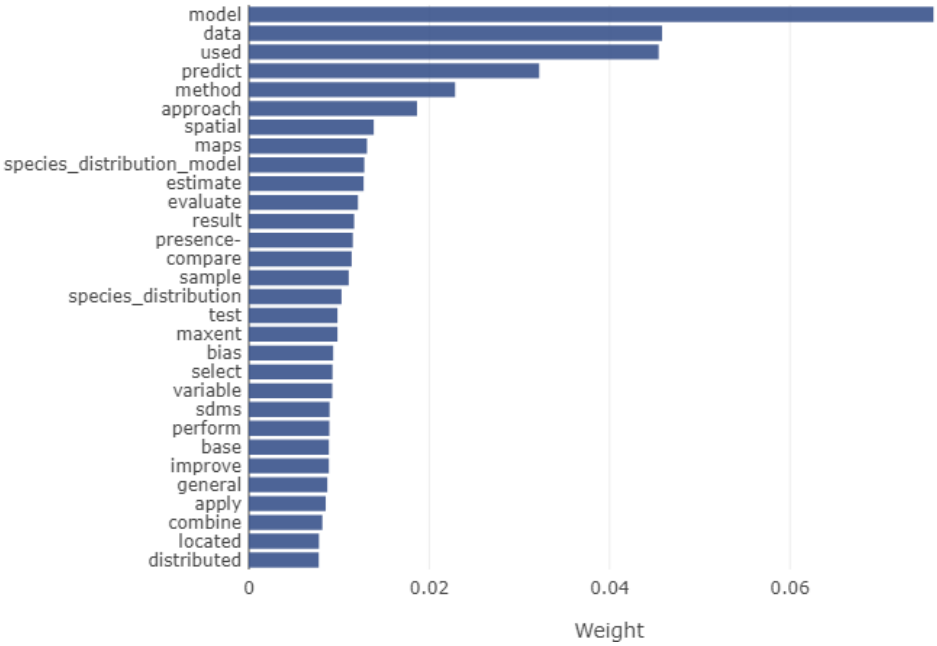

area
distributed
conserve
potential
region
predict
current
high
risk
identify
invasion
climate_change
important
occurrence_records
range
suitable
assess
future
will
include
southern
project
species
northern
suitable_habitat
establish
protect
present
potential_distribution
island
0
0.01
0.02
0.03
0.04
0.05
Weight

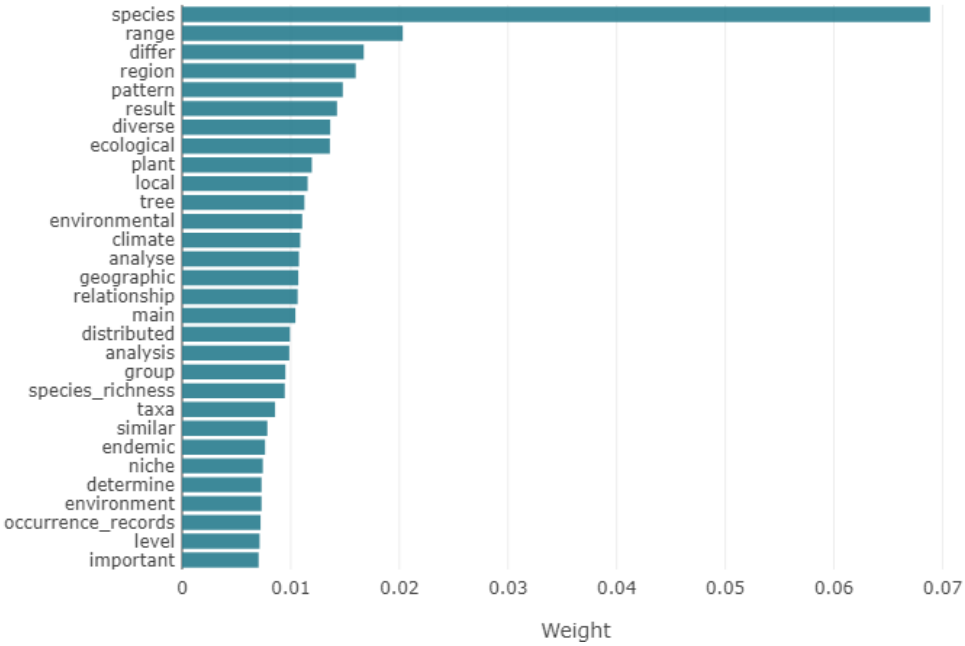

data
species
record
observed
collect
inform
survey
provide
dataset
citizen_science
report
study
monitor
detect
biodiversity
research
include
also
database
available
rare
knowledge
number
represent
source
large
fish
sighting
many
effort
0
0.01
0.02
0.03
0.04
0.05
0.06
Weight

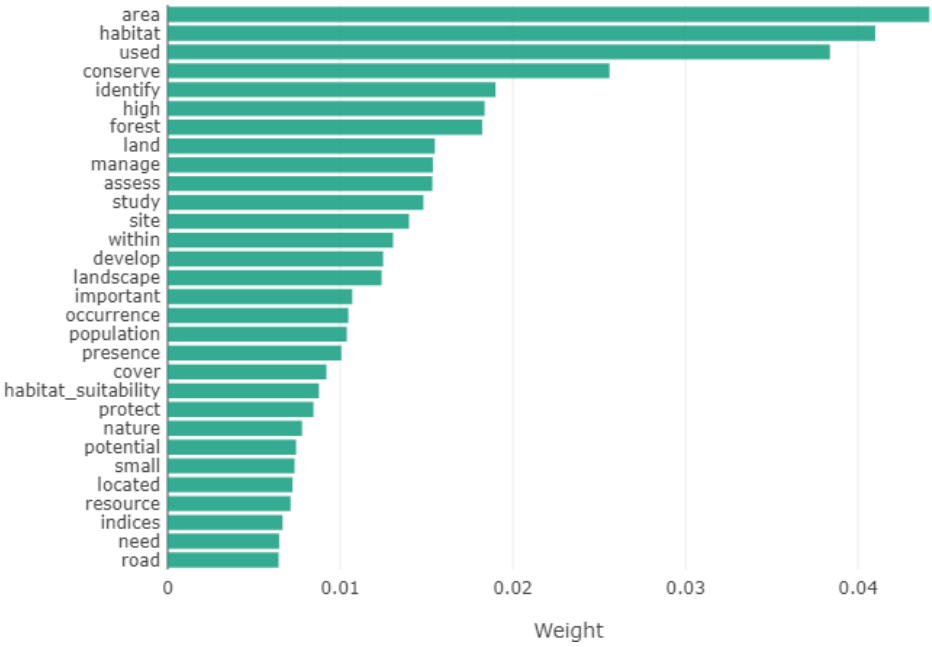
Six clusters

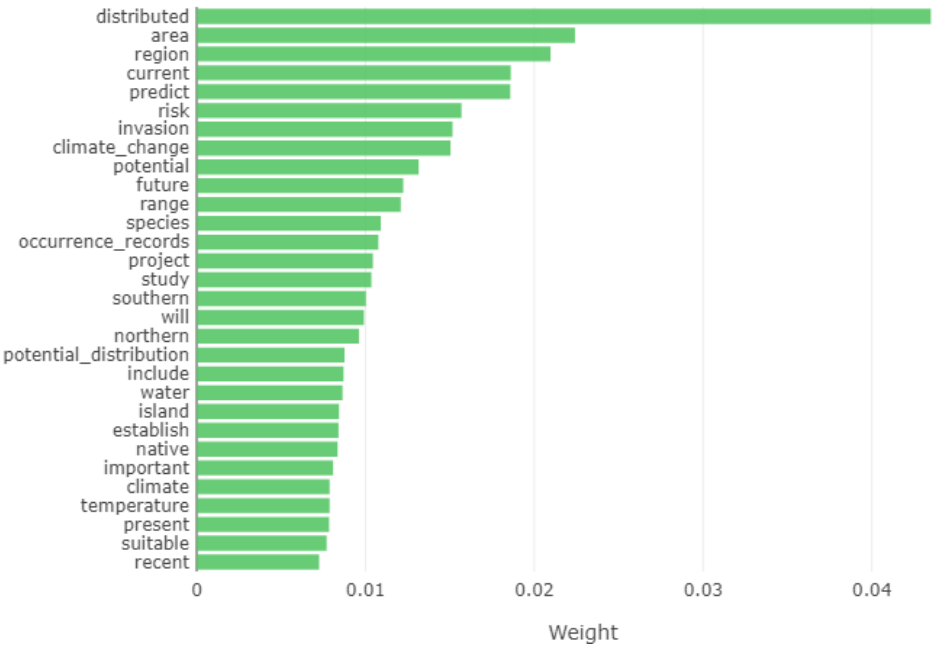


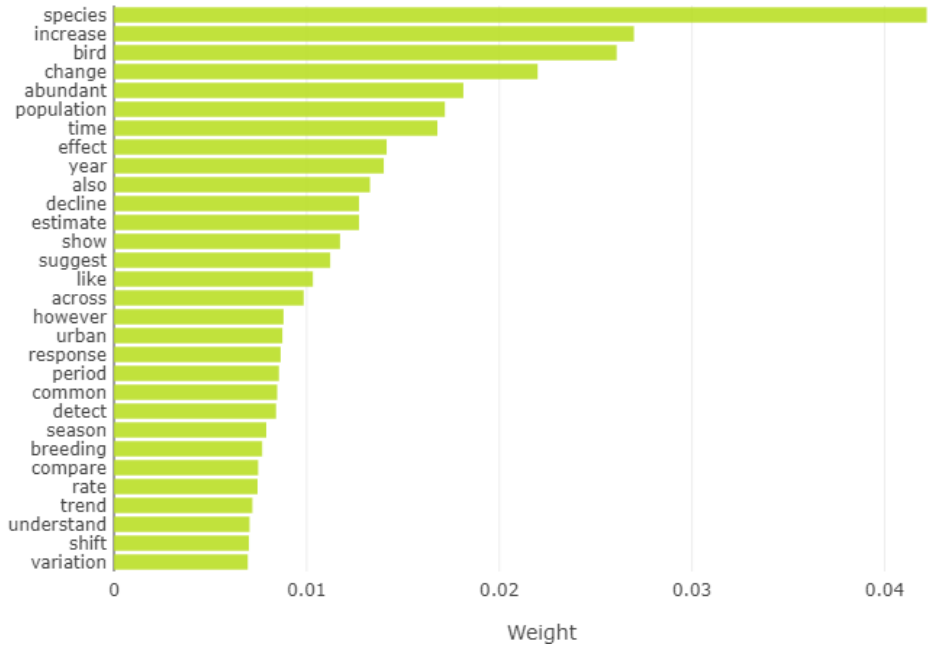

data
record
inform
species
collect
observed
survey
provide
dataset
citizen_science
report
monitor
biodiversity
research
database
study
knowledge
source
include
available
fish
sighting
effort
occurrence
number
addition
large
sample
specimen
also
0
0.02
0.04
0.06
0.08
Weight


model
data
used
predict
method
approach
spatial
maps
species_distribution_model
estimate
evaluate
result
presence-
compare
sample
species_distribution
test
maxent
bias
select
variable
sdms
perform
base
improve
general
apply
combine
located
distributed
0
0.02
0.04
0.06
Weight

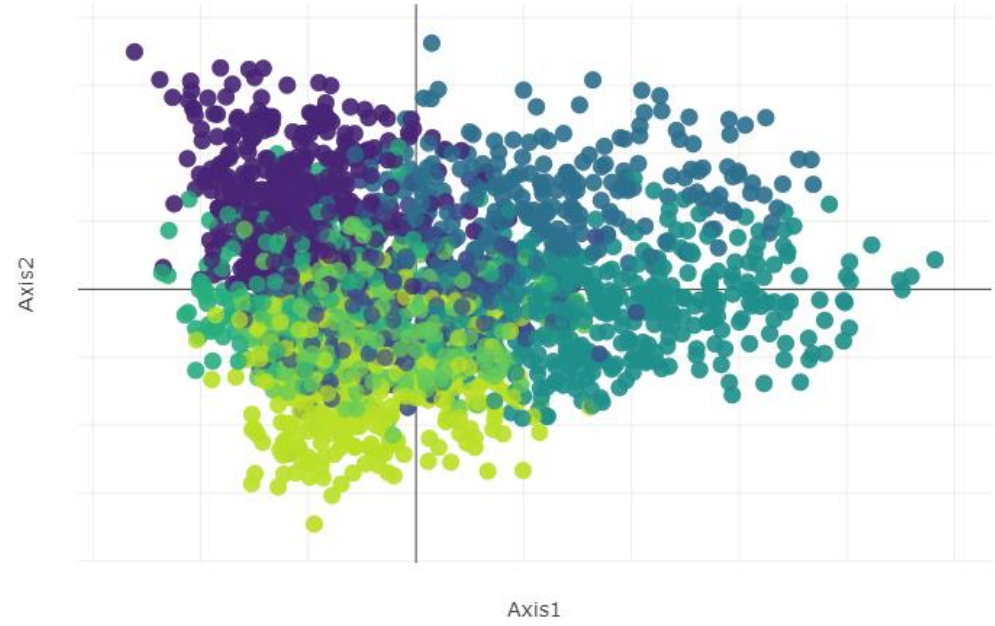

species
range
differ
region
pattern
result
diverse
ecological
plant
local
tree
environmental
climate
analyse
geographic
relationship
main
distributed
analysis
group
species_richness
taxa
similar
endemic
niche
determine
environment
occurrence_records
level
important
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Weight

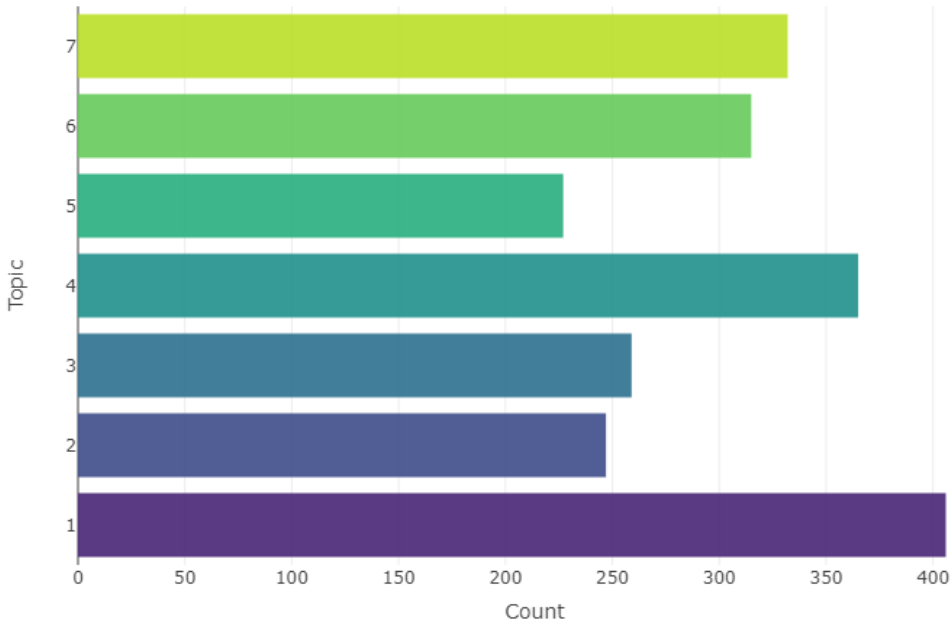

area
habitat
used
conserve
identify
high
forest
land
manage
assess
study
site
within
develop
landscape
important
occurrence
population
presence
cover
habitat_suitability
protect
nature
potential
small
located
resource
indices
need
road
0
0.01
0.02
0.03
0.04
Weight

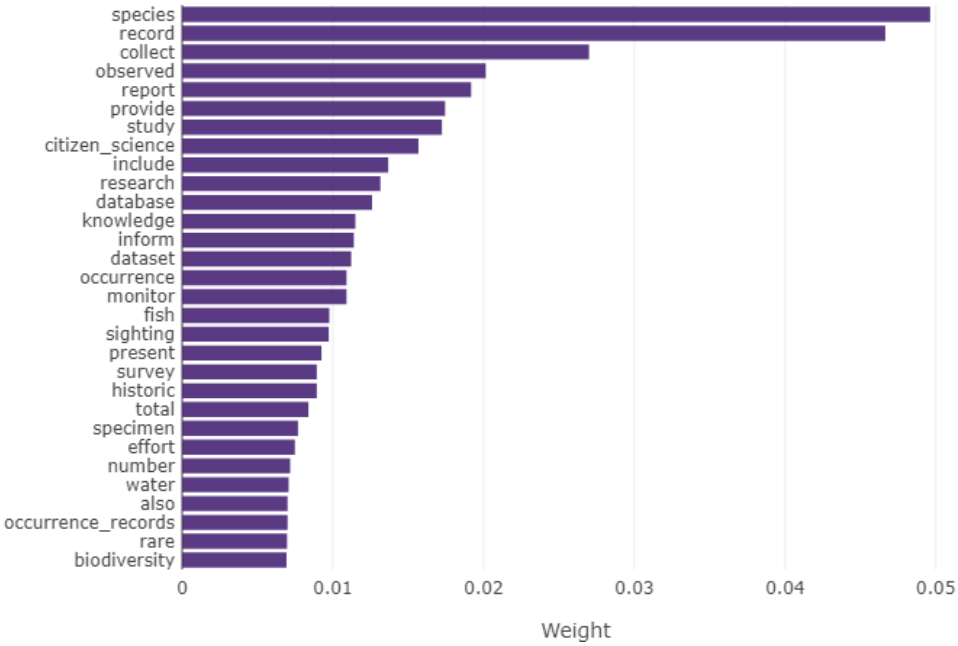

distributed
area
region
current
predict
risk
invasion
climate_change
potential
future
range
species
occurrence_records
project
study
southern
will
northern
potential_distribution
include
water
island
establish
native
important
climate
temperature
present
suitable
recent
0
0.01
0.02
0.03
0.04
Weight

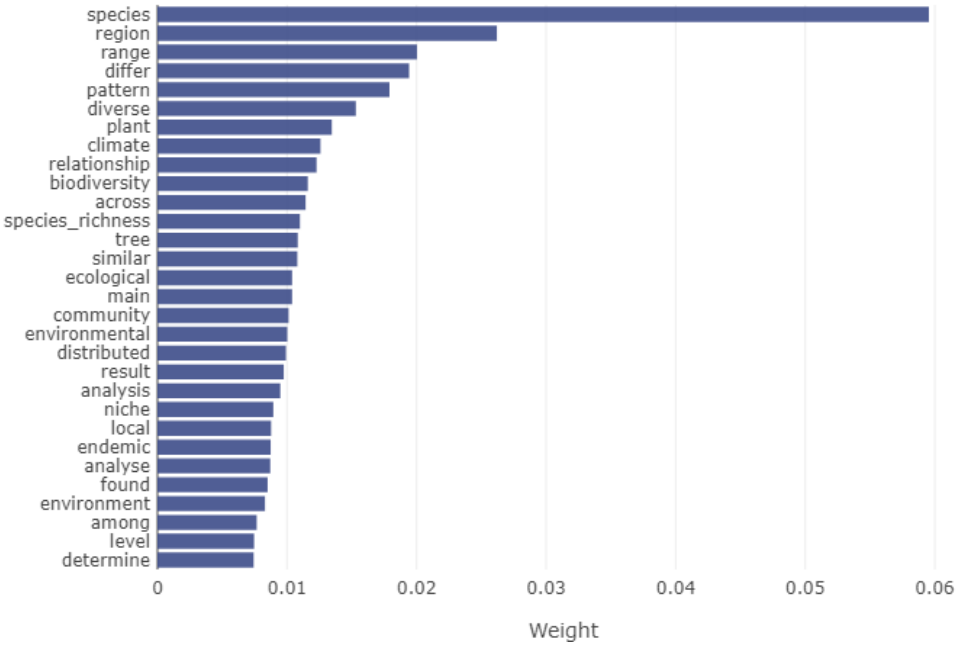

species
increase
bird
change
abundant
population
time
effect
year
also
decline
estimate
show
suggest
like
across
however
urban
response
period
common
detect
season
breeding
compare
rate
trend
understand
shift
variation
0
0.01
0.02
0.03
0.04
Weight

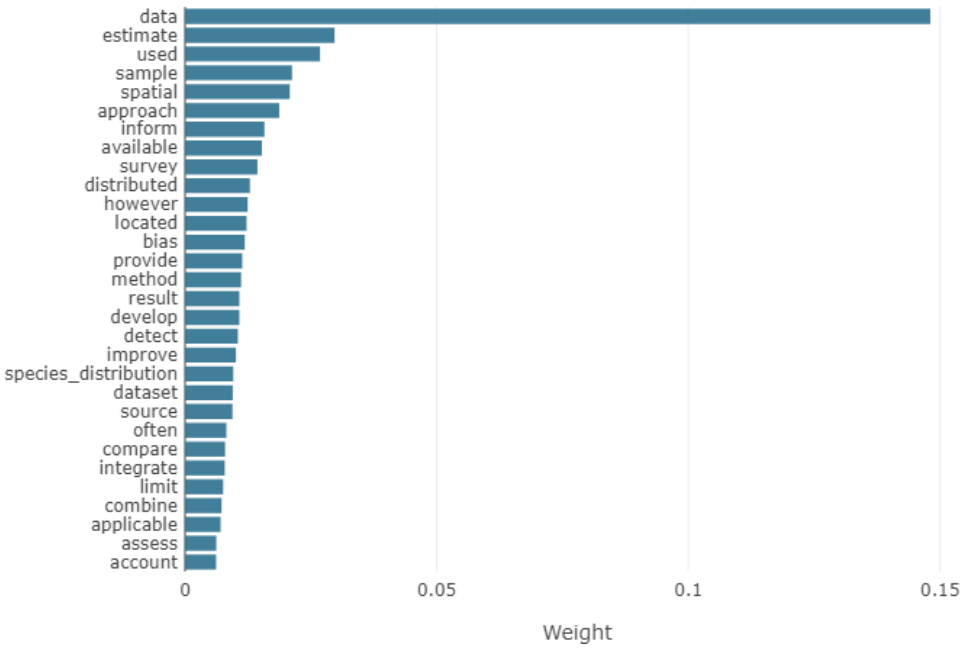
Seven clusters

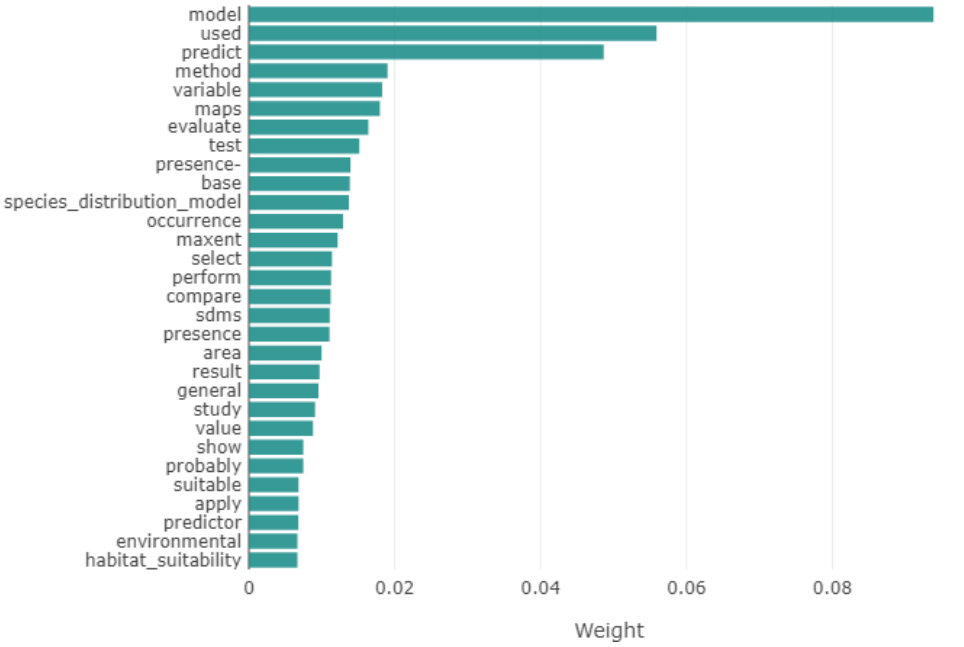


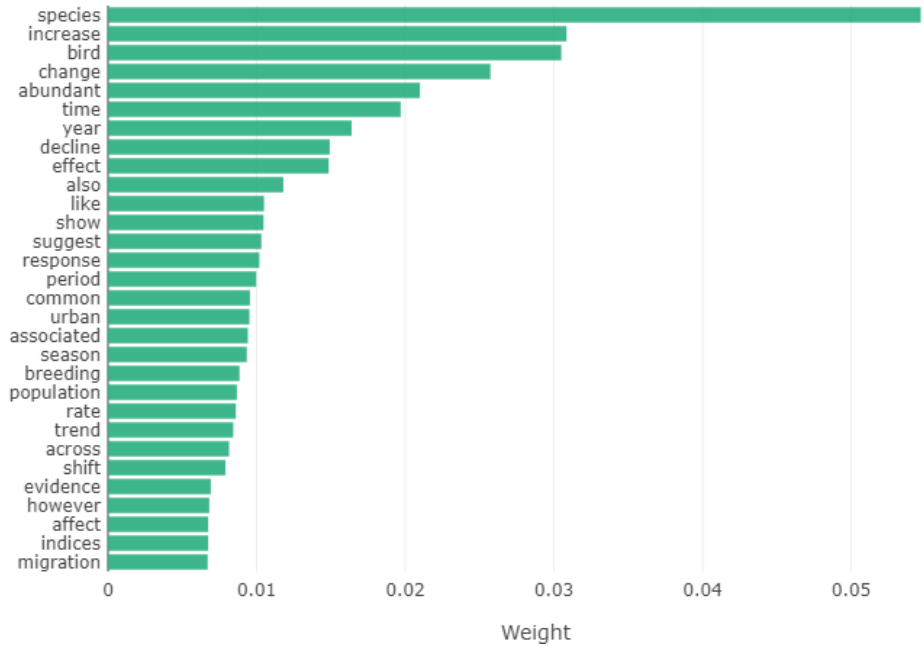


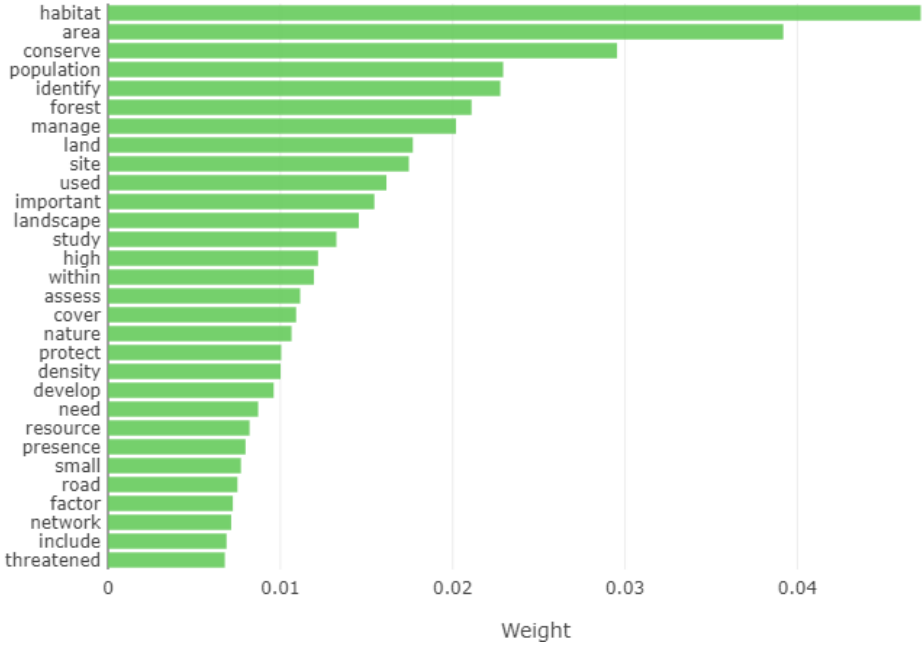

species
record
collect
observed
report
provide
study
citizen_science
include
research
database
knowledge
inform
dataset
occurrence
monitor
fish
sighting
present
survey
historic
total
specimen
effort
number
water
also
occurrence_records
rare
biodiversity
0
0.01
0.02
0.03
0.04
0.05
Weight

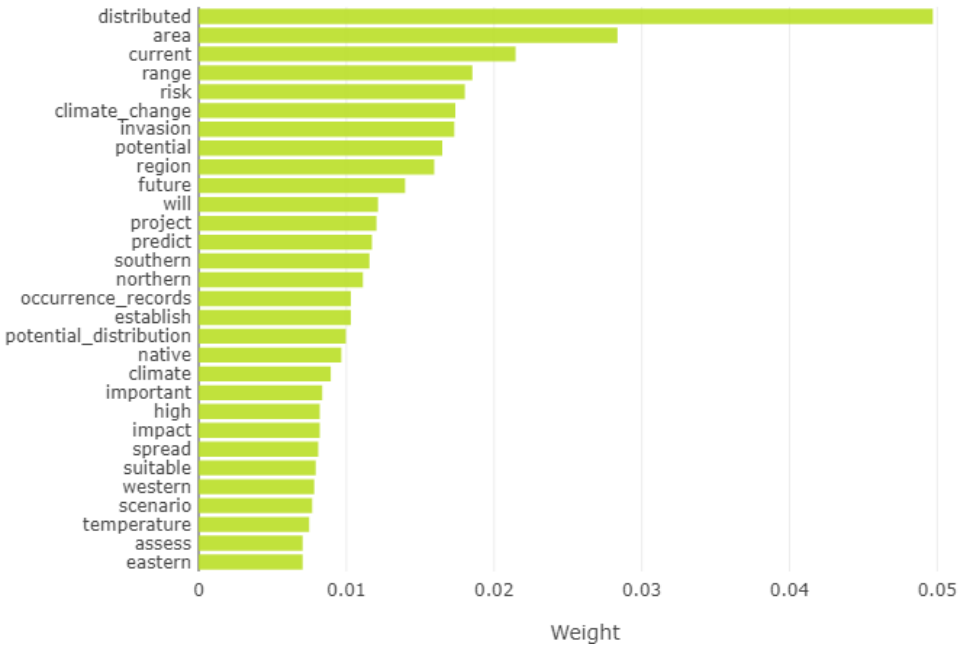

species
region
range
differ
pattern
diverse
plant
climate
relationship
biodiversity
across
species_richness
tree
similar
ecological
main
community
environmental
distributed
result
analysis
niche
local
endemic
analyse
found
environment
among
level
determine
0
0.01
0.02
0.03
0.04
0.05
0.06
Weight


data
estimate
used
sample
spatial
approach
inform
available
survey
distributed
however
located
bias
provide
method
result
develop
detect
improve
species_distribution
dataset
source
often
compare
integrate
limit
combine
applicable
assess
account
0
0.05
0.1
0.15
Weight

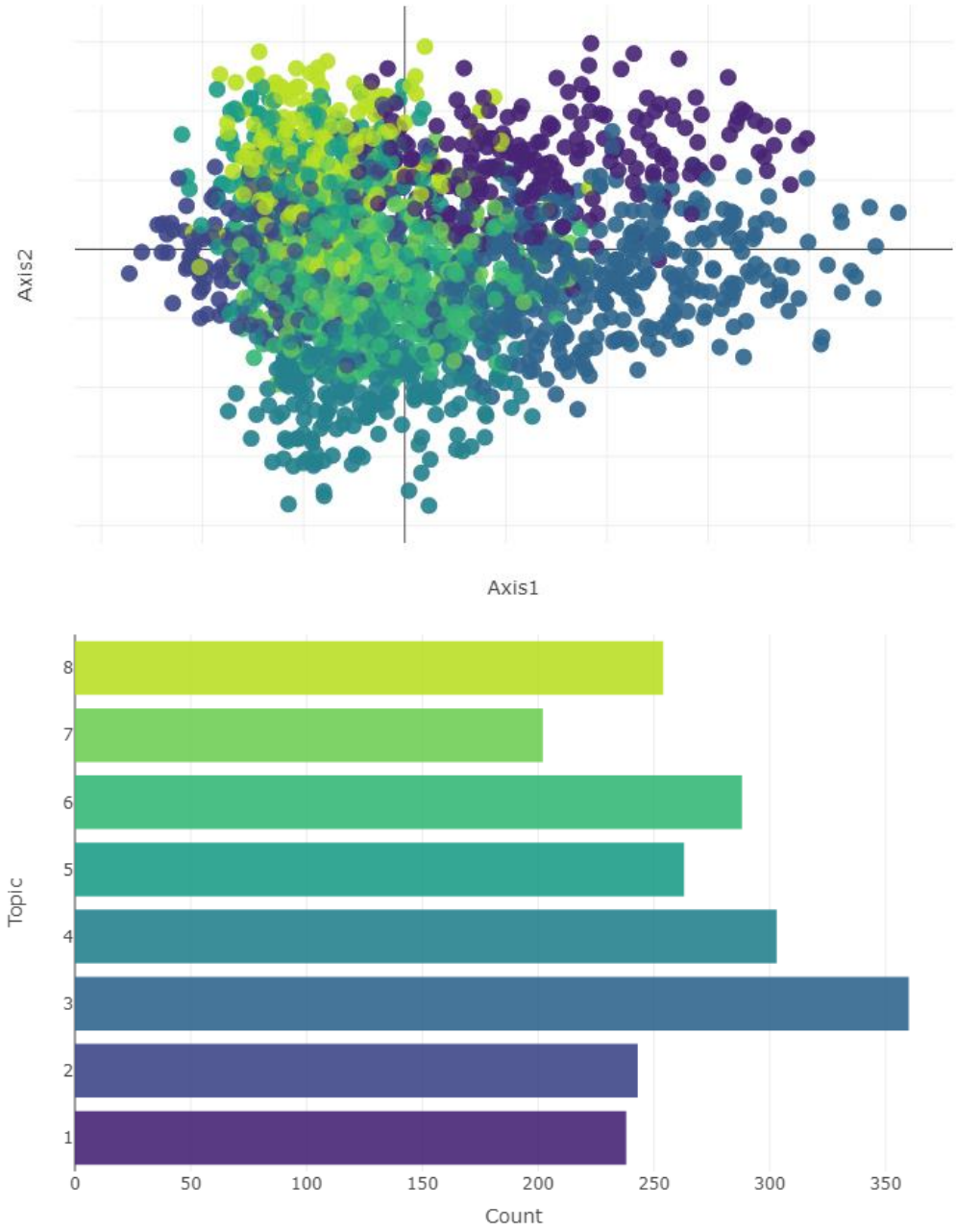

model
used
predict
method
variable
maps
evaluate
test
presence-
base
species_distribution_model
occurrence
maxent
select
perform
compare
sdms
presence
area
result
general
study
value
show
probably
suitable
apply
predictor
environmental
habitat_suitability
0
0.02
0.04
0.06
0.08
Weight

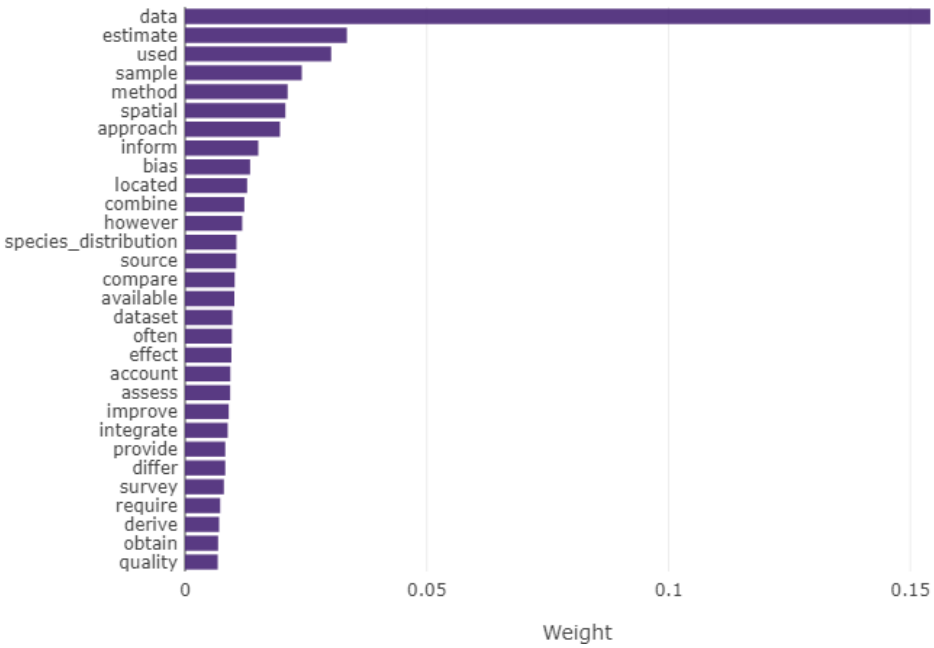

species
increase
bird
change
abundant
time
year
decline
effect
also
like
show
suggest
response
period
common
urban
associated
season
breeding
population
rate
trend
across
shift
evidence
however
affect
indices
migration
0
0.01
0.02
0.03
0.04
0.05
Weight

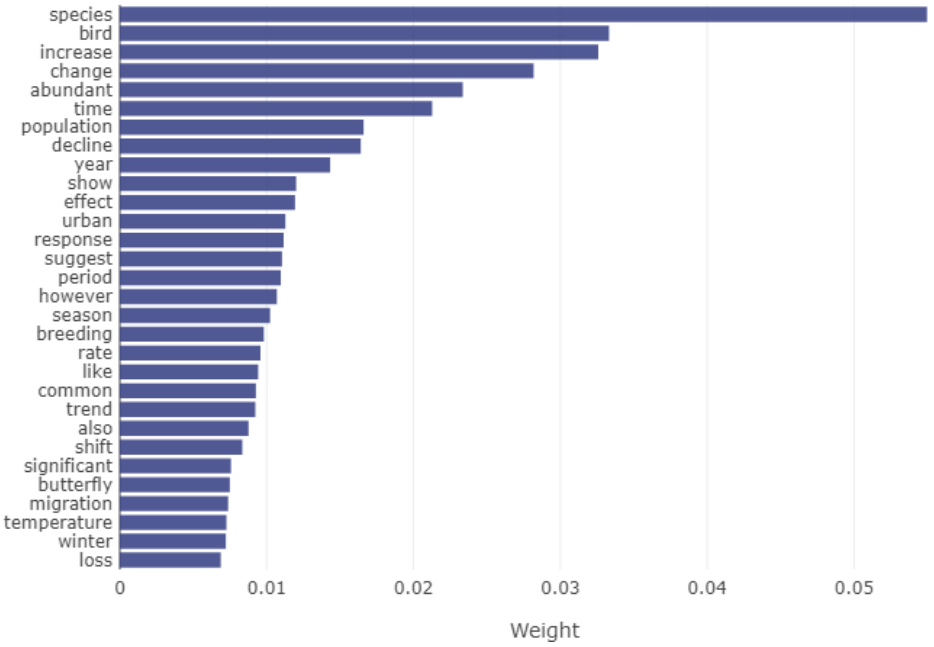

habitat
area
conserve
population
identify
forest
manage
land
site
used
important
landscape
study
high
within
assess
cover
nature
protect
density
develop
need
resource
presence
small
road
factor
network
include
threatened
0
0.01
0.02
0.03
0.04
Weight

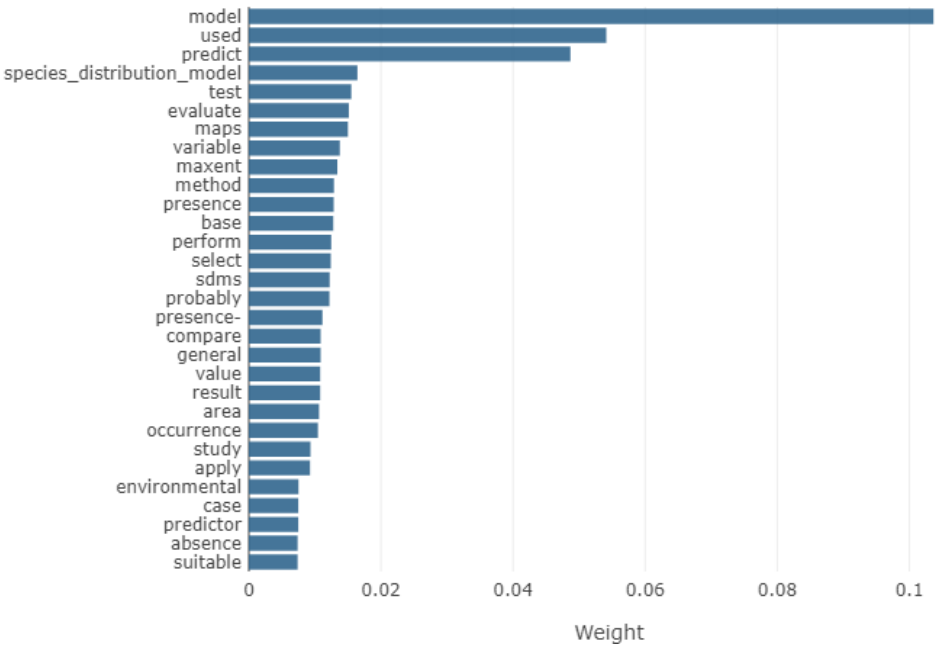

distributed
area
current
range
risk
climate_change
invasion
potential
region
future
will
project
predict
southern
northern
occurrence_records
establish
potential_distribution
native
climate
important
high
impact
spread
suitable
western
scenario
temperature
assess
eastern
0
0.01
0.02
0.03
0.04
0.05
Weight

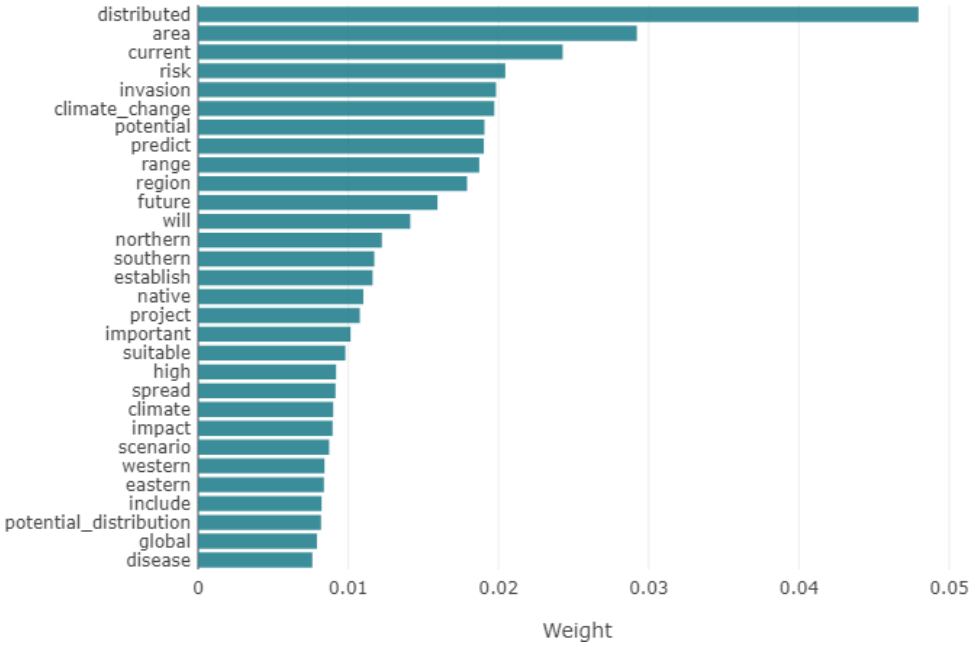
Eight clusters

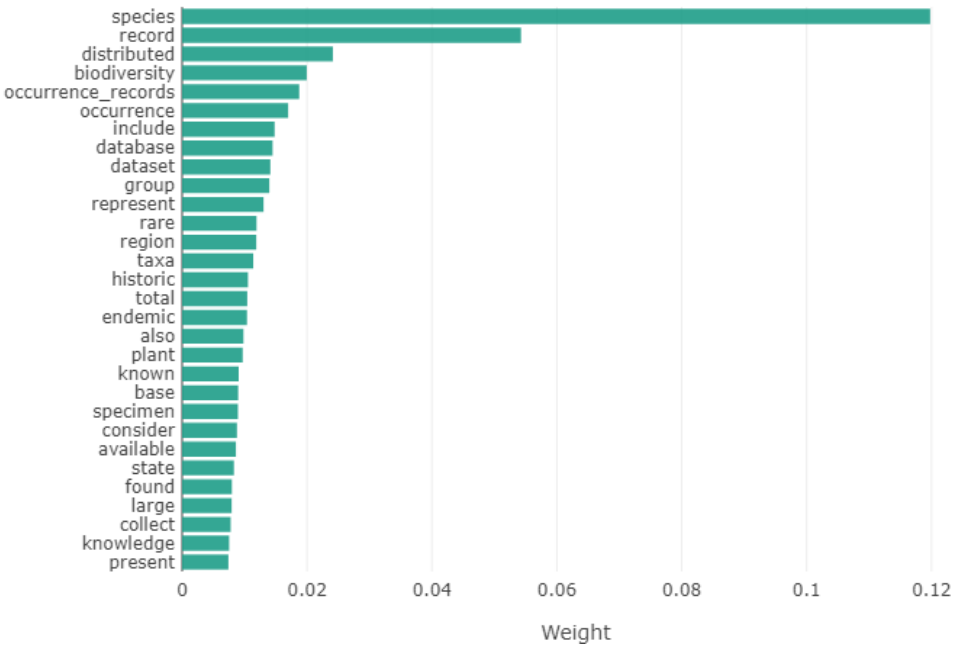


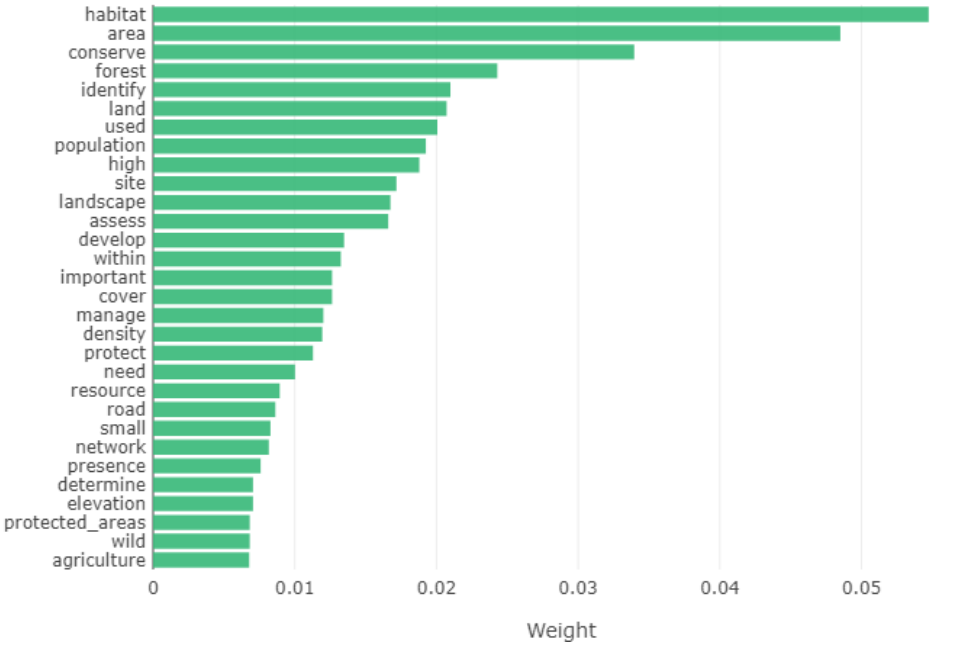

data
estimate
used
sample
method
spatial
approach
inform
bias
located
combine
however
species_distribution
source
compare
available
dataset
often
effect
account
assess
improve
integrate
provide
differ
survey
require
derive
obtain
quality
0
0.05
0.1
0.15
Weight

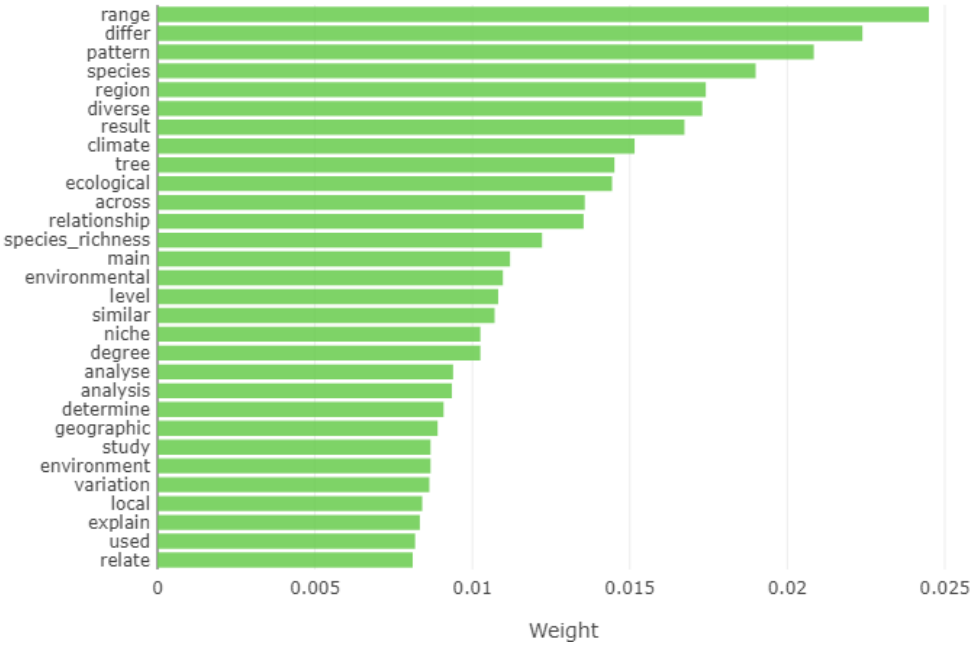

species
bird
increase
change
abundant
time
population
decline
year
show
effect
urban
response
suggest
period
however
season
breeding
rate
like
common
trend
also
shift
significant
butterfly
migration
temperature
winter
loss
0
0.01
0.02
0.03
0.04
0.05
Weight

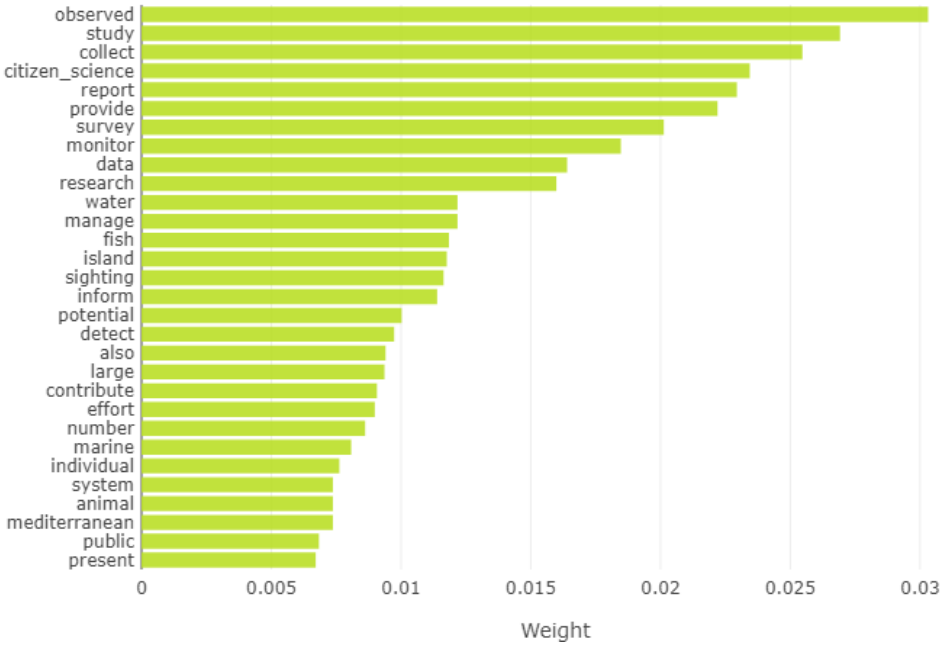

model
used
predict
species_distribution_model
test
evaluate
maps
variable
maxent
method
presence
base
perform
select
sdms
probably
presence-
compare
general
value
result
area
occurrence
study
apply
environmental
case
predictor
absence
suitable
0
0.02
0.04
0.06
0.08
0.1
Weight


distributed
area
current
risk
invasion
climate_change
potential
predict
range
region
future
will
northern
southern
establish
native
project
important
suitable
high
spread
climate
impact
scenario
western
eastern
include
potential_distribution
global
disease
0
0.01
0.02
0.03
0.04
0.05
Weight


species
record
distributed
biodiversity
occurrence_records
occurrence
include
database
dataset
group
represent
rare
region
taxa
historic
total
endemic
also
plant
known
base
specimen
consider
available
state
found
large
collect
knowledge
present
0
0.02
0.04
0.06
0.08
0.1
0.12
Weight


habitat
area
conserve
forest
identify
land
used
population
high
site
landscape
assess
develop
within
important
cover
manage
density
protect
need
resource
road
small
network
presence
determine
elevation
protected_areas
wild
agriculture
0
0.01
0.02
0.03
0.04
0.05
Weight


range
differ
pattern
species
region
diverse
result
climate
tree
ecological
across
relationship
species_richness
main
environmental
level
similar
niche
degree
analyse
analysis
determine
geographic
study
environment
variation
local
explain
used
relate
0
0.005
0.01
0.015
0.02
0.025
Weight


observed
study
collect
citizen_science
report
provide
survey
monitor
data
research
water
manage
fish
island
sighting
inform
potential
detect
also
large
contribute
effort
number
marine
individual
system
animal
mediterranean
public
present
0
0.005
0.01
0.015
0.02
0.025
0.03
Weight

S3. Data sheet categories used to categorize the set of 300 papers read in full.

All variable categories, except for study region, were not mutually exclusive; that is, an article could be coded with as many variable responses as applicable.

Variable category	Variable (True/False Response)
Topic categories mentioned in abstract [True/false]	Invasive species
	Land use change
	Climate change
	Overexploitation
	Pollution
	Other conservation issues
	Other basic ecology topics
	New methods development
	Comparing multiple presence-only approaches
	Comparing presence-only with more structured approaches
	Testing methodological choices within one presence-only approach
	Testing new technology for analyzing/reporting presence-only data
Taxa [True/false]	Bird
	Mammal
	Amphibian/reptile
	Fish
	Invertebrate
	Virus/bacteria/similar
	Plant/similar
Study system [True/false]	Terrestrial
	Marine
	Freshwater
Study region [True/false]	Africa
	Asia
	Europe
	Latin America
	North America
	Oceania
	Oceans
	Polar regions
	Multiple/global
Author region [True/false][1]	Africa

	Asia Europe Latin America North America Oceania
Study scale [True/false]	Local (upper size limit defined as municipality) Regional (upper size limit defined as large state/province and/or small nation) Large (defined as large national to continental scale) Global (defined as multiple continents)
Sample size [True/false]	1-10 11-100 101-1,000 1,001-10,000 10,001-100,000 100,001-1,000,000 > 1,000,000 Not described
Sampling design [True/false]	Explicitly described as opportunistic Semi-structured sampling design Structured presence/absence data Not described
Direct data source [Number of each type of source, unless otherwise noted]	Original data Large openly accessible database Small openly accessible database Literature Social media Unpublished data/personal communication Private organization/nonprofit Government agency Museum/herbarium/collections [For open databases] Name of database [open-ended response field] [For open databases] Is open database still available? [True/false]
Original data source [True/false]	Citizen science
Data availability [True/false, unless otherwise noted]	All data shared after publication – in an open database All data shared after publication – other method Location/format of shared data [open-ended response field] Is shared data still accessible?

Analysis approach [True/false]	Report of occurrence
	Spatial summary statistics
	Analysis of user trends
	Species distribution/ecological niche modeling
	Occupancy modeling
	List length analysis
	Species richness/diversity measures
	Phenology
	Population dynamics/demographic modeling
	Multivariate analyses
Other analysis information [True/false]	Comparison with more structured analysis types
	Integration with more structured data types
	Presence-only data used to evaluate a different type of analysis
	Presence-only data used to design a different type of analysis
	Biases associated with presence-only data discussed

[1] Study and author region categories were derived from the GBIF Regions[4].

S4. Data collected from the set of 300 articles.

Data are available at `https://doi.org/10.17605/OSF.IO/JUEQC`.

S5. Ten most cited articles and most commonly cited references among included articles.

Table S5a. The ten most cited articles from within the articles included in our review.

Article	**Times cited**
Phillips et al. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling.	7546
Phillips and Dudík 2008. Modeling of species distributions with Maxent. Ecography.	3063

Elith et al. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions.	2658
Pearson et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography.	1540
Hernandez et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography.	1258
Phillips et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications.	1251
Merow et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography.	1129
Anderson et al. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models Ecological Modelling.	712
Engler et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology.	576
Pearson et al. 2006. Model-based uncertainty in species range prediction. Journal of Biogeography.	556

Table S5b. The most common references cited by articles included in our review. These references are not necessarily within the set of articles included in our review.

Article	**Times referenced**
Phillips et al. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling.	695

Elith et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography.	509
Hijmans et al. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology.	429
Phillips and Dudík 2008. Modeling of species distributions with Maxent. Ecography.	347
Fielding and Bell 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation.	322
Elith et al. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions.	306
Guisan and Zimmermann 2000. Predictive habitat distribution models in ecology. Ecological Modelling.	291
Phillips et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications.	289
Elith and Leathwick 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics.	263
Guisan and Thuiller 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters.	261

S6. Openly accessible databases used by articles in the set of 300.

Asterisk indicates databases considered 'large' for the purpose of this review.

Database	Times used
* Global Biodiversity Information Facility (GBIF)	37
* eBird	9
* Atlas of Living Australia	8
* iNaturalist	8
* Tropicos	8
* OBIS	4
* speciesLink	4
Butterflies for the New Millenium	3
* FishBase	3
Victorian Biodiversity Atlas	3
Birdlife Australia	2
Biodiversity Information Serving Our Nation (US) (BISON)	2
BugGuide.net	2
Butterfly Conservation	2
Chinese Virtual Herbarium	2
Dutch National Database Flora and Fauna	2
EDDMaps	2
* iDigBio	2
Joint Nature Conservancy Council Seabird Censuses	2
ManisNet.org	2
National Specimen Information Infrastructure (China)	2
SEINet Portal Network	2
Swedish Lifewatch	2
Taiwan Roadkill Observation Network	2
UK Biological Records Centre	2
VertNet	2
WikiAves	2
AK Libellen NRW	1
AquaNIS	1
ArtDatabanken (Swedish Species Observation System)	1
Artsdatabanken (Norwegian Biodiversity Information Centre)	1
Atlas of New South Wales Wildlife	1
Aves de Chile	1
Base de Datos sobre Scarabaeidae (BANDASCA)	1
Basking Shark Watch (UK Marine Conservation Society)	1
Biodiversity Databank of Catalonia	1

BioObs	1
Bird Conservation Society of Thailand (BCST)	1
BirdLife Finland Tiira database	1
Birdlife International	1
BOLD (Barcode of Life)	1
British Dragonfly Society Recording Scheme	1
British Trust for Ornithology	1
Butterflies and Moths of North America	1
CalOdes	1
CardObs	1
Centre for Agriculture and Biosciences International (CABI)	1
Centre of Environmental Data and Recording (CEDaR) (North Ireland)	1
Centre Suisse de la Faune	1
cloudbirders.com	1
COL (National Colombian Herbarium of the Instituto de Ciencias Naturales)	1
Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO)	1
Database for Ecosystems and Ecosystem Service Zoning in China	1
Datenbank Artenschutzkartierung	1
Données d'Observations pour la Reconnaissance et l'Identification de la faune et la flore Subaquatiques (DORIS)	1
Dutch Butterfly Monitoring Scheme	1
Dutch Dragonfly Monitoring Scheme	1
eButterfly	1
EPPO Global Database (European and Mediterranean Plant Protection Organization)	1
EUFORGEN (European Forest Genetic Resources Programme)	1
European Environment Agency (http://eunis.eea.europa.eu)	1
falterfunde.de (science4you)	1
Faune-Aquitaine	1
Flora of Cyprus	1
Flora-On	1
Flotrop	1
Global Ant Biodiversity Informatics (GABI)	1
Global Mammal Parasite Database	1
HOLOS Ecoinformatics Engine	1
http://magicicada.org/	1
http://mammiferimarini.unipv.it/ Strandings Database	1

https://www.geocetus.it/ Stranding Information System	1
https://www.ornitho.at/	1
https://www.ornitho.ch/	1
https://www.ornitho.it/	1
Influenza Research Database (FluDB)	1
INPN Espèces	1
Insect Database (Finnish Museum of Natural History)	1
insecte.org	1
iSeahorse	1
JABOT (Rio de Janeiro Botanical Garden)	1
Jaguar GIS (http://www.savethejaguar.org)	1
JellyWatch (http://www.jellywatch.org)	1
JSTOR	1
LANDFIRE reference data base 2010, v1.2.0	1
Malaysian Nature Society Bird i-Witness database	1
Massachusetts Audubon Butterfly Atlas	1
MosquitoMap	1
Natagora	1
National Biodiversity Data Centre (Ireland)	1
National Indigenous Vegetation Survey Database (New Zealand)	1
National Institute of Invasive Species Science (NIISS) database (US)	1
NeoTropTree	1
New Zealand Herpetofauna Database	1
North American Breeding Bird Survey	1
OBIS-SEAMAP	1
Observadores del Mar	1
Odonata Central	1
PERSEUS (Policy-oriented marine Environmental Research in the Southern European Seas)	1
REBIOMA (Réseau de la Biodiversité de Madagascar)	1
Red de Observadores de Libélulas de Andalucía	1
Redmap	1
Reef Life Survey	1
Seaquest Southwest, Cornwall Wildlife Trust	1
SIG-Ivoire	1
Société française d'Odonatologie	1
SOMBASE (Southern Ocean Mollusc Database)	1
SWEMP (Southwest Exotic Mapping Program)	1
The Database on Taxonomy of Drosophilidae	1
Tokyo Butterfly Monitoring	1
UK National Biodiversity Network	1

University of British Columbia E-fauna	1
USGS Nonindigenous Aquatic Species database (US)	1
VectorMap	1
West African Vegetation database of the Senckenberg Research Institute	1
www.naturbeobachtung.at	1
xeno-canto	1

S7. Bibliography of the subset of 300 articles that were randomly selected from the full set of 2151 articles to be read in full and coded for analysis.

Abood F, Murphy RJ. 2006. World distribution of Minthea rugicollis (coleoptera: lyctidae). Journal of Tropical Forest Science 18: 250-254.

Acosta LE. 2014. Bioclimatic profile and potential distribution of the Mesopotamian harvestman Discocyrtus testudineus (Holmberg, 1876) (Opiliones, Gonyleptidae). Zootaxa 3821: 301.

Adhikari D, Tiwary R, Singh PP, Upadhaya K, Singh B, Haridasan KE, Bhatt BB, Chettri A, Barik SK. 2019. Ecological niche modeling as a cumulative environmental impact assessment tool for biodiversity assessment and conservation planning: A case study of critically endangered plant Lagerstroemia minuticarpa in the Indian Eastern Himalaya. Journal of Environmental Management 243: 299–307.

Alahmed AM, Naeem M, Kheir SM, Sallam MF. Ecological Distribution Modeling of Two Malaria Mosquito Vectors Using Geographical Information System in Al-Baha Province, Kingdom of Saudi Arabia. 10.

Albuquerque F, Benito B, Rodriguez MÁM, Gray C. 2018. Potential changes in the distribution of Carnegiea gigantea under future scenarios. PeerJ 6: e5623.

Alimi TO, Fuller DO, Qualls WA, Herrera SV, Arevalo-Herrera M, Quinones ML, Lacerda MVG, Beier JC. 2015. Predicting potential ranges of primary malaria vectors and malaria in northern South America based on projected changes in climate, land cover and human population. Parasites & Vectors 8: 431.

Alkhamis MA, Arruda AG, Vilalta C, Morrison RB, Perez AM. 2018. Surveillance of porcine reproductive and respiratory syndrome virus in the United States using risk mapping and species distribution modeling. Preventive Veterinary Medicine 150: 135–142.

Allouche O, Steinitz O, Rotem D, Rosenfeld A, Kadmon R. 2008. Incorporating distance constraints into species distribution models. Journal of Applied Ecology 45: 599–609.

Anderson RP, Lew D, Peterson AT. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162: 211–232.

Armstrong AJ, Armstrong AO, Bennett MB, McGregor F, Abrantes KG, Barnett A, Richardson AJ, Townsend KA, Dudgeon CL. 2020. The geographic distribution of reef and oceanic manta rays (Mobula alfredi and Mobula birostris) in Australian coastal waters. Journal of Fish Biology 96: 835–840.

Ashrafzadeh MR, Naghipour AA, Haidarian M, Kusza S, Pilliod DS. 2019. Effects of climate change on habitat and connectivity for populations of a vulnerable, endemic salamander in Iran. Global Ecology and Conservation 19: e00637.

Aubry KB, Lewis JC. 2003. Extirpation and reintroduction of fishers (Martes pennanti) in Oregon: implications for their conservation in the Pacific states. Biological Conservation 114: 79–90.

Ay J-S, Guillemot J, Martin-StPaul N, Doyen L, Leadley P. 2017. The economics of land use reveals a selection bias in tree species distribution models: Selection bias in species distribution models. Global Ecology and Biogeography 26: 65–77.

Aylesworth L, Phoonsawat R, Suvanachai P, Vincent ACJ. 2017. Generating spatial data for marine conservation and management. Biodiversity and Conservation 26: 383–399.

Ballesteros-Mejia L, Kitching IJ, Jetz W, Beck J. 2017. Putting insects on the map: near-global variation in sphingid moth richness along spatial and environmental gradients. Ecography 40: 698–708.

Baltensperger AP, Morton JM, Huettmann F. 2017. Expansion of American marten (Martes americana) distribution in response to climate and landscape change on the Kenai Peninsula, Alaska. Journal of Mammalogy 98: 703–714.

Bariche M. 2018. First confirmed record of the white-spotted puffer Arothron hispidus (Linnaeus, 1758) in the Mediterranean Sea. BioInvasions Records 7: 433–436.

Beasley-Hall PG, Lee TRC, Rose HA, Lo N. 2018. Multiple abiotic factors correlate with parallel evolution in Australian soil burrowing cockroaches. Journal of Biogeography 45: 1515–1528.

Bedriñana-Romano L, Hucke-Gaete R, Viddi FA, Morales J, Williams R, Ashe E, Garcés-Vargas J, Torres-Florez JP, Ruiz J. 2018. Integrating multiple data sources for assessing blue whale abundance and distribution in Chilean Northern Patagonia. Diversity and Distributions 24: 991–1004.

Belkhiria J, Hijmans RJ, Boyce W, Crossley BM, Martínez-López B. 2018. Identification of high risk areas for avian influenza outbreaks in California using disease distribution models. PLOS ONE 13: e0190824.

Beneš J, Konvička M, Vrabec V, Zámečník J. 2003. Do the sibling species of small whites, Leptidea sinapis and L. reali (Lepidoptera, Pieridae) differ in habitat preferences? Biologia 58: 943-951.

Benito X, Trobajo R, Ibáñez C. 2014. Modelling Habitat Distribution of Mediterranean Coastal Wetlands: The Ebro Delta as Case Study. Wetlands 34: 775–785.

Berzaghi F, Engel JE, Plumptre AJ, Mugabe H, Kujirakwinja D, Ayebare S, Bates JM. 2018. Comparative niche modeling of two bush-shrikes (Laniarius) and the conservation of mid-elevation Afromontane forests of the Albertine Rift. The Condor 120: 803–814.

Bezuijen MR. 2000. The occurrence of the flat-headed cat Prionailurus planiceps in south-east Sumatra. Oryx 34: 222-226.

Birkmanis CA, Partridge JC, Simmons LW, Heupel MR, Sequeira AMM. 2020. Shark conservation hindered by lack of habitat protection. Global Ecology and Conservation 21: e00862.

Boakes EH, Fuller RA, McGowan PJK, Mace GM. 2016. Uncertainty in identifying local extinctions: the distribution of missing data and its effects on biodiversity measures. Biology Letters 12: 20150824.

Bocksberger G, Schnitzler J, Chatelain C, Daget P, Janssen T, Schmidt M, Thiombiano A, Zizka G. 2016. Climate and the distribution of grasses in West Africa. Journal of Vegetation Science 27: 306–317.

Bolliger J, Kienast F, Soliva R, Rutherford G. 2007. Spatial sensitivity of species habitat patterns to scenarios of land use change (Switzerland). Landscape Ecology 22: 773–789.

Bonnet-Lebrun A -S., Karamanlidis AA, de Gabriel Hernando M, Renner I, Gimenez O. 2020. Identifying priority conservation areas for a recovering brown bear population in Greece using citizen science data. Animal Conservation 23: 83–93.

Bosch S, Tyberghein L, Deneudt K, Hernandez F, De Clerck O. 2018. In search of relevant predictors for marine species distribution modelling using the MarineSPEED benchmark dataset. Diversity and Distributions 24: 144–157.

Botella C, Joly A, Bonnet P, Monestiez P, Munoz F. 2018. Species distribution modeling based on the automated identification of citizen observations. Applications in Plant Sciences 6: e1029.

Bottin M, Peyre G, Vargas C, Raz L, Richardson JE, Sanchez A. 2020. Phytosociological data and herbarium collections show congruent large-scale patterns but differ in their local descriptions of community composition. Journal of Vegetation Science 31: 208–219.

Botts EA, Erasmus BFN, Alexander GJ. 2015. Observed range dynamics of South African amphibians under conditions of global change: Changes in Amphibian Distribution. Austral Ecology 40: 309–317.

Bradsworth N, White JG, Isaac B, Cooke R. 2017. Species distribution models derived from citizen science data predict the fine scale movements of owls in an urbanizing landscape. Biological Conservation 213: 27–35.

Bradter U, Mair L, Jönsson M, Knape J, Singer A, Snäll T. 2018. Can opportunistically collected Citizen Science data fill a data gap for habitat suitability models

of less common species? Methods in Ecology and Evolution 9: 1667–1678.
Brambilla M, Falco R, Negri I. 2012. A spatially explicit assessment of within-season changes in environmental suitability for farmland birds along an altitudinal gradient: Within-season changes in habitat suitability in birds. Animal Conservation 15: 638–647.
Brambilla M, Ficetola GF. 2012. Species distribution models as a tool to estimate reproductive parameters: a case study with a passerine bird species: Distribution models and reproductive parameters. Journal of Animal Ecology 81: 781–787.
Brambilla M, Scridel D, Bazzi G, Ilahiane L, Iemma A, Pedrini P, Bassi E, Bionda R, Marchesi L, Genero F, Teufelbauer N, Probst R, Vrezec A, Kmecl P, Mihelič T, Bogliani G, Schmid H, Assandri G, Pontarini R, Braunisch V, Arlettaz R, Chamberlain D. 2020. Species interactions and climate change: How the disruption of species co-occurrence will impact on an avian forest guild. Global Change Biology 26: 1212–1224.
Braz AG, de Viveiros Grelle CE, de Souza Lima Figueiredo M, Weber M de M. 2020. Interspecific competition constrains local abundance in highly suitable areas. Ecography 43: 1560–1570.
Breed GA, Stichter S, Crone EE. 2013. Climate-driven changes in northeastern US butterfly communities. Nature Climate Change 3: 142–145.
Briscoe D, Hiatt S, Lewison R, Hines E. 2014. Modeling habitat and bycatch risk for dugongs in Sabah, Malaysia. Endangered Species Research 24: 237–247.
Brown DM, Sieswerda PL, Parsons ECM. 2019. Potential encounters between humpback whales (Megaptera novaeangliae) and vessels in the New York Bight apex, USA. Marine Policy 106: 103527.
Brown JL, Yoder AD. 2015. Shifting ranges and conservation challenges for lemurs in the face of climate change. Ecology and Evolution 5: 1131–1142.
Brown S, Clarke M, Clarke R. 2009. Fire is a key element in the landscape-scale habitat requirements and global population status of a threatened bird: The Mallee Emu-wren (Stipiturus mallee). Biological Conservation 142: 432–445.
Byers JE, McDowell WG, Dodd SR, Haynie RS, Pintor LM, Wilde SB. 2013. Climate and pH Predict the Potential Range of the Invasive Apple Snail (Pomacea insularum) in the Southeastern United States. PLoS ONE 8: e56812.
Bystriakova N, Peregrym M, Dragicevic S. 2015. Effect of environment on distributions of rock ferns in the Mediterranean climate: The case of the genus Asplenium in Montenegro. Flora - Morphology, Distribution, Functional Ecology of Plants 215: 84–91.
Camacho C. 2016. Birding trip reports as a data source for monitoring rare species. Animal Conservation 19: 430–435.
Cao B, Bai C, Xue Y, Yang J, Gao P, Liang H, Zhang L, Che L, Wang J, Xu J, Duan C, Mao M, Li G. 2020. Wetlands rise and fall: Six endangered wetland species showed different patterns of habitat shift under future climate change. Science of The Total Environment 731: 138518.

Carota C, Nava CR, Ghiglione C, Schiaparelli S. 2017. A Bayesian semiparametric GLMM for historical and newly collected presence-only data: An application to species richness of Ross Sea Mollusca: A Bayesian semiparametric GLMM for presence-only data. Environmetrics 28: e2462.

Carrascal LM, Aragón P, Palomino D, Lobo JM. 2015. Predicting regional densities from bird occurrence data: validation and effects of species traits in a Macaronesian Island. Diversity and Distributions 21: 1284–1294.

Castuera-Oliveira L, Oliveira-Filho AT de, Eisenlohr PV. 2020. Emerging hotspots of tree richness in Brazil. Acta Botanica Brasilica 34: 117–134.

Cerasoli F, Thuiller W, Guéguen M, Renaud J, D'Alessandro P, Biondi M. 2020. The role of climate and biotic factors in shaping current distributions and potential future shifts of European Neocrepidodera (Coleoptera, Chrysomelidae). Insect Conservation and Diversity 13: 47–62.

César de Sá N, Marchante H, Marchante E, Cabral JA, Honrado JP, Vicente JR. 2019. Can citizen science data guide the surveillance of invasive plants? A model-based test with Acacia trees in Portugal. Biological Invasions 21: 2127–2141.

Changeux T, Blazy C, Ruitton S. 2020. The use of citizen science for marine biodiversity surveys: from species identification to ecologically relevant observations. Hydrobiologia 847: 27–43.

Chapman D, Pescott OL, Roy HE, Tanner R. 2019. Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. Journal of Biogeography 46: 1029–1040.

Chen H, Chen L, Albright TP. 2007. Predicting the potential distribution of invasive exotic species using GIS and information-theoretic approaches: A case of ragweed (Ambrosia artemisiifolia L.) distribution in China. Chinese Science Bulletin 52: 1223–1230.

Cheyne SM, Mohamed A, Hearn AJ, Ross J, Samejima H, Heydon M, Augeri DM, van Berkel T, Boonratana R, Fredriksson G, Hon J, Marshall J, Macdonald DW, Belant JL, Kramer-Schadt S, Wilting A. 2016. Predicted distribution of the otter civet Cynogale bennettii (Mammalia: Carnivora: Viverridae) on Borneo. The Raffles Bulletin of Zoology 33: 126-131.

Christodoulou CS, Griffiths GH, Vogiatzakis IN. 2018. Using threatened plant species to identify conservation gaps and opportunities on the island of Cyprus. Biodiversity and Conservation 27: 2837–2858.

Chyn K, Lin T-E, Chen Y-K, Chen C-Y, Fitzgerald LA. 2019. The magnitude of roadkill in Taiwan: Patterns and consequences revealed by citizen science. Biological Conservation 237: 317–326.

Číhal L, Kaláb O, Plášek V. 2017. Modeling the distribution of rare and interesting moss species of the family Orthotrichaceae (Bryophyta) in Tajikistan and Kyrgyzstan. Acta Societatis Botanicorum Poloniae 86: 3543.

Clements GR, Rayan DM, Aziz SA, Kawanishi K, Traeholt C, Magintan D, Yazi MFA, Tingley R. 2012. Predicting the distribution of the Asian tapir in

Peninsular Malaysia using maximum entropy modeling. Integrative Zoology 7: 400–406.

Coleman T, Mentch L, Fink D, La Sorte FA, Winkler DW, Hooker G, Hochachka WM. 2020. Statistical inference on tree swallow migrations with random forests. Journal of the Royal Statistical Society: Series C (Applied Statistics) 69: 973–989.

Colli-Silva M, Pirani JR. 2020. Estimating bioregions and undercollected areas in South America by revisiting Byttnerioideae, Helicteroideae and Sterculioideae (Malvaceae) occurrence data. Flora 271: 151688.

Collins SD, Abbott JC, McIntyre NE. 2017. Quantifying the degree of bias from using county-scale data in species distribution modeling: Can increasing sample size or using county-averaged environmental data reduce distributional overprediction? Ecology and Evolution 7: 6012–6022.

Condet M, Dulau-Drouot V. 2016. Habitat selection of two island-associated dolphin species from the south-west Indian Ocean. Continental Shelf Research 125: 18–27.

Cooley JR, Arguedas N, Bonaros E, Bunker G, Chiswell SM, DeGiovine A, Edwards M, Hassanieh D, Haji D, Knox J, Kritsky G, Mills C, Mozgai D, Troutman R, Zyla J, Hasegawa H, Sota T, Yoshimura J, Simon C. 2018. The periodical cicada four-year acceleration hypothesis revisited and the polyphyletic nature of Brood V, including an updated crowd-source enhanced map (Hemiptera: Cicadidae: Magicicada). PeerJ 6: e5282.

Coro G, Pagano P, Ellenbroek A. 2013. Combining simulated expert knowledge with Neural Networks to produce Ecological Niche Models for Latimeria chalumnae. Ecological Modelling 268: 55–63.

Coron C, Calenge C, Giraud C, Julliard R. 2018. Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. Environmental and Ecological Statistics 25: 71–93.

Costa H, Medeiros V, Azevedo EB, Silva L. 2013. Evaluating ecological-niche factor analysis as a modelling tool for environmental weed management in island systems. Weed Research 53: 221–230.

Costa H, Ponte NB, Azevedo EB, Gil A. 2015. Fuzzy set theory for predicting the potential distribution and cost-effective monitoring of invasive species. Ecological Modelling 316: 122–132.

Couturier T, Besnard A, Bertolero A, Bosc V, Astruc G, Cheylan M. 2014. Factors determining the abundance and occurrence of Hermann's tortoise Testudo hermanni in France and Spain: Fire regime and landscape changes as the main drivers. Biological Conservation 170: 177–187.

Coxen CL, Frey JK, Carleton SA, Collins DP. 2017. Species distribution models for a migratory bird based on citizen science and satellite tracking data. Global Ecology and Conservation 11: 298–311.

Crawford PHC, Hoagland BW. 2010. Using species distribution models to guide conservation at the state level: the endangered American burying beetle

(Nicrophorus americanus) in Oklahoma. Journal of Insect Conservation 14: 511–521.

Cruz-Cárdenas G, López-Mata L, Villaseñor JL, Ortiz E. 2014. Potential species distribution modeling and the use of principal component analysis as predictor variables. Revista Mexicana de Biodiversidad 85: 189–199.

Dallas T, Huang S, Nunn C, Park AW, Drake JM. 2017. Estimating parasite host range. Proceedings of the Royal Society B: Biological Sciences 284: 20171250.

De Wysiecki AM, Sánchez-Carnero N, Irigoyen AJ, Milessi AC, Colonello JH, Bovcon ND, Cortés F, Barbini SA, Cedrola PV, Coller NM, Jaureguizar AJ. 2020. Using temporally explicit habitat suitability models to infer the migratory pattern of a large mobile shark. Canadian Journal of Fisheries and Aquatic Sciences 77: 1529–1539.

Deb JC, Rahman HMT, Roy A. 2016. Freshwater Swamp Forest Trees of Bangladesh Face Extinction Risk from Climate Change. Wetlands 36: 323–334.

Deidun A, Castriota L, Falautano M, Maggio T. 2017a. Yet another angelfish species for the Mediterranean – the first record of Holacanthus africanus Cadenat, 1951 from Maltese waters, central Mediterranean. BioInvasions Records 6: 373–376.

Deidun A, Sciberras J, Sciberras A, Gauci A, Balistreri P, Salvatore A, Piraino S. 2017b. The first record of the white-spotted Australian jellyfish Phyllorhiza punctata von Lendenfeld, 1884 from Maltese waters (western Mediterranean) and from the Ionian coast of Italy. BioInvasions Records 6: 119–124.

Dennis EB, Morgan BJT, Freeman SN, Ridout MS, Brereton TM, Fox R, Powney GD, Roy DB. 2017. Efficient occupancy model-fitting for extensive citizen-science data. PLOS ONE 12: e0174433.

Dias D, Baringo Fonseca C, Correa L, Soto N, Portela A, Juarez K, Tumolo Neto RJ, Ferro M, Gonçalves J, Junior J. 2017. Repatriation Data: More than two million species occurrence records added to the Brazilian Biodiversity Information Facility Repository (SiBBr). Biodiversity Data Journal 5: e12012.

Dimson M, Lynch SC, Gillespie TW. 2019. Using biased sampling data to model the distribution of invasive shot-hole borers in California. Biological Invasions 21: 2693–2712.

Ding F, Ma T, Hao M, Wang Q, Chen S, Wang D, Huang L, Zhang X, Jiang D. 2020. Mapping Worldwide Environmental Suitability for Artemisia annua L. Sustainability 12: 1309.

Dorazio RM. 2012. Predicting the Geographic Distribution of a Species from Presence-Only Data Subject to Detection Errors. Biometrics 68: 1303–1312.

Drake JM, Randin C, Guisan A. 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43: 424–432.

Duan R-Y, Kong X-Q, Huang M-Y, Fan W-Y, Wang Z-G. 2014. The Predictive Performance and Stability of Six Species Distribution Models. PLoS ONE 9: e112764.

Dwyer RG, Carpenter-Bundhoo L, Franklin CE, Campbell HA. 2016. Using citizen-collected wildlife sightings to predict traffic strike hot spots for threatened species: a case study on the southern cassowary. Journal of Applied Ecology 53: 973–982.

Dyer RJ, Gillings S, Pywell RF, Fox R, Roy DB, Oliver TH. 2017. Developing a biodiversity-based indicator for large-scale environmental assessment: a case study of proposed shale gas extraction sites in Britain. Journal of Applied Ecology 54: 872–882.

El-Gabbas A, Dormann CF. 2018. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography 41: 1161–1172.

Etherington TR, Ward AI, Smith GC, Pietravalle S, Wilson GJ. 2009. Using the Mahalanobis distance statistic with unplanned presence-only survey data for biogeographical models of species distribution and abundance: a case study of badger setts. Journal of Biogeography 36: 845–853.

Falk BG, Snow RW, Reed RN. 2016. Prospects and Limitations of Citizen Science in Invasive Species Management: A Case Study with Burmese Pythons in Everglades National Park. Southeastern Naturalist 15: 89–102.

Ferreira S, Tierno de Figueroa JM, Martins F, Verissimo J, Quaglietta L, Grosso-Silva JM, Lopes P, Sousa P, Paupério J, Fonseca N, Beja P. 2020. The InBIO Barcoding Initiative Database: contribution to the knowledge on DNA barcodes of Iberian Plecoptera. Biodiversity Data Journal 8: e55137.

Flaherty M, Lawton C. 2019. The regional demise of a non-native invasive species: the decline of grey squirrels in Ireland. Biological Invasions 21: 2401–2416.

Forero-Medina G, Cárdenas-Arévalo G, Castaño-Mora OV. Habitat modeling of Dahl's toad-headed turtle (Mesoclemmys dahli) in Colombia. Herpetological Conservation and Biology 7: 313-322.

Frey JK, Calkins MT. 2014. Snow cover and riparian habitat determine the distribution of the short-tailed weasel (Mustela erminea) at its southern range limits in arid western North America. 78: 45-56.

Froese JG, Smith CS, Durr PA, McAlpine CA, van Klinken RD. 2017. Modelling seasonal habitat suitability for wide-ranging species: Invasive wild pigs in northern Australia. PLOS ONE 12: e0177018.

Gamliel I, Buba Y, Guy-Haim T, Garval T, Willette D, Rilov G, Belmaker J. 2020. Incorporating physiology into species distribution models moderates the projected impact of warming on selected Mediterranean marine species. Ecography 43: 1090–1106.

Gebrewahid Y, Abrehe S, Meresa E, Eyasu G, Abay K, Gebreab G, Kidanemariam K, Adissu G, Abreha G, Darcha G. 2020. Current and future predicting potential areas of Oxytenanthera abyssinica (A. Richard) using MaxEnt model under climate change in Northern Ethiopia. Ecological Processes 9: 6.

Gerstner BE, Kass JM, Kays R, Helgen KM, Anderson RP. 2018. Revised distributional estimates for the recently discovered olinguito (Bassaricyon neblina),

with comments on natural and taxonomic history. Journal of Mammalogy 99: 321–332.
Ghisbain G, Williams PH, Michez D, Branstetter MG, Rasmont P. 2020. Contribution to the knowledge of the bumblebee fauna of Afghanistan (Hymenoptera, Apidae, Bombus Latreille). ZooKeys 973: 69–87.
Gibson L, Barrett B, Burbidge A. 2007. Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot: Uncertain absences in habitat modelling of a rare bird. Diversity and Distributions 13: 704–713.
Girardello M, Chapman A, Dennis R, Kaila L, Borges PAV, Santangeli A. 2019. Gaps in butterfly inventory data: A global analysis. Biological Conservation 236: 289–295.
Goberville E, Hautekèete N-C, Kirby RR, Piquot Y, Luczak C, Beaugrand G. 2016. Climate change and the ash dieback crisis. Scientific Reports 6: 35303.
Golding N, Purse BV. 2016. Fast and flexible Bayesian species distribution modelling using Gaussian processes. Methods in Ecology and Evolution 7: 598–608.
Goldstein EA, Lawton C, Sheehy E, Butler F. 2014. Locating species range frontiers: a cost and efficiency comparison of citizen science and hair-tube survey methods for use in tracking an invasive squirrel. Wildlife Research 41: 64.
Gómez-Martínez MA, Klem D, Rojas-Soto O, González-García F, MacGregor-Fors I. 2019. Window strikes: bird collisions in a Neotropical green city. Urban Ecosystems 22: 699–708.
Gooliaff T, Weir RD, Hodges KE. 2018. Estimating bobcat and Canada lynx distributions in British Columbia: Estimating Bobcat and Lynx Distributions. The Journal of Wildlife Management 82: 810–820.
Gottwald J, Appelhans T, Adorf F, Hillen J, Nauss T. 2017. High-Resolution MaxEnt Modelling of Habitat Suitability for Maternity Colonies of the Barbastelle Bat Barbastella barbastellus (Schreber, 1774) in Rhineland-Palatinate, Germany. Acta Chiropterologica 19: 389–398.
Griffin SC, Taper ML, Hoffman R, Mills LS. 2010. Ranking Mahalanobis Distance Models for Predictions of Occupancy From Presence-Only Data. Journal of Wildlife Management 74: 1112–1121.
Guevara L, León-Paniagua L. 2019. How to survive a glaciation: the challenge of estimating biologically realistic potential distributions under freezing conditions. Ecography 42: 1237–1245.
Guisan A, Zimmermann NE, Elith J, Graham CH, Phillips S, Peterson AT. 2007. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? Ecological Monographs 77: 615–630.
Hanberry BB, He HS, Palik BJ. 2012. Pseudoabsence Generation Strategies for Species Distribution Models. PLoS ONE 7: e44486.
Hann C, Stelle L, Szabo A, Torres L. 2018. Obstacles and Opportunities of Using a Mobile App for Marine Mammal Research. ISPRS International Journal of

Geo-Information 7: 169.
Haque MdM, Nipperess DA, Gallagher RV, Beaumont LJ. 2017. How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. Austral Ecology 42: 690–699.
Harvey DS, Platenberg RJ. Predicting habitat use from opportunistic observations: a case study of the Virgin Islands tree boa (Epicrates granti). Herpetological Journal 19: 111-118.
Hefley TJ, Hooten MB. 2015. On the existence of maximum likelihood estimates for presence-only data. Methods in Ecology and Evolution 6: 648–655.
Hengl T, Sierdsema H, Radović A, Dilo A. 2009. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. Ecological Modelling 220: 3499–3511.
Heringer G, Almeida TE, Dittrich VA de O, Salino A. 2020. Assessing the effectiveness of protected areas for the conservation of ferns and lycophytes in the Brazilian state of Minas Gerais. Journal for Nature Conservation 53: 125775.
Hernandez PA, Graham CH, Master LL, Albert DL. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29: 773–785.
Herrera Campo BV, Hyman G, Bellotti A. 2011. Threats to cassava production: known and potential geographic distribution of four key biotic constraints. Food Security 3: 329–345.
Hertzog LR, Besnard A, Jay-Robert P. 2014. Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. Diversity and Distributions 20: 1403–1413.
Higa M, Yamaura Y, Koizumi I, Yabuhara Y, Senzaki M, Ono S. 2015. Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. Diversity and Distributions 21: 46–54.
Hill NJ, Tobin AJ, Reside AE, Pepperell JG, Bridge TCL. 2016. Dynamic habitat suitability modelling reveals rapid poleward distribution shift in a mobile apex predator. Global Change Biology 22: 1086–1096.
Hughes AC, Satasook C, Bates PJJ, Soisook P, Sritongchuay T, Jones G, Bumrungsri S. 2010. Echolocation Call Analysis and Presence-Only Modelling as Conservation Monitoring Tools for Rhinolophoid Bats in Thailand. Acta Chiropterologica 12: 311–327.
Isaac B, White J, Ierodiaconou D, Cooke R. 2013. Response of a cryptic apex predator to a complete urban to forest gradient. Wildlife Research 40: 427.
Iturbide M, Bedia J, Gutiérrez JM. 2018. Background sampling and transferability of species distribution model ensembles under climate change. Global and Planetary Change 166: 19–29.
Jackson MM, Gergel SE, Martin K. 2015. Effects of Climate Change on Habitat Availability and Configuration for an Endemic Coastal Alpine Bird. PLOS ONE 10: e0142110.

Jarnevich CS, Talbert M, Morisette J, Aldridge C, Brown CS, Kumar S, Manier D, Talbert C, Holcombe T. 2017. Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: An example with background selection. Ecological Modelling 363: 48–56.

Jaźwa M, Jedrzejczak E, Klichowska E, Pliszko A. Predicting the potential distribution area of Solidago ×niederederi (Asteraceae). Turkish Journal of Botany 42: 51-56.

Jezkova T, Wiens JJ. 2018. Testing the role of climate in speciation: New methods and applications to squamate reptiles (lizards and snakes). Molecular Ecology 27: 2754–2769.

Jiménez-Valverde A, Decae AE, Arnedo MA. 2011. Environmental suitability of new reported localities of the funnelweb spider Macrothele calpeiana: an assessment using potential distribution modelling with presence-only techniques: Environmental suitability of new localities of Macrothele calpeiana. Journal of Biogeography 38: 1213–1223.

Jiménez-Valverde A, Ortuño VM, Lobo JM. 2007. Exploring the distribution of Sterocorax Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula: Distribution of Sterocorax species in the Iberian Peninsula. Journal of Biogeography 34: 1426–1438.

Jinga P, Ashley MV. 2019. Climate change threatens some miombo tree species of sub-Saharan Africa. Flora 257: 151421.

Jinga P, Palagi J. 2020. Dry and wet miombo woodlands of south-central Africa respond differently to climate change. Environmental Monitoring and Assessment 192: 372.

Jones A, Bruce E, Davies KP, Blewitt M, Sheehan S. 2019. Assessing potential environmental influences on killer whale (Orcinus orca) distribution patterns in the Bremer Canyon, south-west Australia. Australian Geographer 50: 381–405.

Joyce M, Barnes MD, Possingham HP, Van Rensburg BJ. 2018. Understanding avian assemblage change within anthropogenic environments using citizen science data. Landscape and Urban Planning 179: 81–89.

Justine J-L, Winsor L, Gey D, Gros P, Thévenot J. 2018. Giant worms chez moi! Hammerhead flatworms (Platyhelminthes, Geoplanidae, Bipalium spp., Diversibipalium spp.) in metropolitan France and overseas French territories. PeerJ 6: e4672.

Kaminski DJ, Poole KE, Clark KB, Harms TM. 2020. Predicting landscape-scale summer resource selection for the northern long-eared bat (Myotis septentrionalis) in Iowa. Journal of Mammalogy 101: 172–186.

Kanagaraj R, Wiegand T, Kramer-Schadt S, Anwar M, Goyal SP. 2011. Assessing habitat suitability for tiger in the fragmented Terai Arc Landscape of India and Nepal. Ecography 34: 970–981.

Keil P, Wilson AM, Jetz W. 2014. Uncertainty, priors, autocorrelation and disparate data in downscaling of species distributions. Diversity and Distributions 20: 797–812.

Kienberger K, Prieto L. 2018. The jellyfish Rhizostoma luteum (Quoy & Gaimard, 1827): not such a rare species after all. Marine Biodiversity 48: 1455–1462.

Kojola I, Heikkinen S, Holmala K. 2018. Balancing costs and confidence: volunteer-provided point observations, GPS telemetry and the genetic monitoring of Finland's wolves. Mammal Research 63: 415–423.

Kolanowska M. 2013. Glacial refugia and migration routes of the Neotropical genus Trizeuxis (Orchidaceae). Acta Societatis Botanicorum Poloniae 82: 225-230.

Kramer-Schadt S, Niedballa J, Pilgrim JD, Schröder B, Lindenborn J, Reinfelder V, Stillfried M, Heckmann I, Scharf AK, Augeri DM, Cheyne SM, Hearn AJ, Ross J, Macdonald DW, Mathai J, Eaton J, Marshall AJ, Semiadi G, Rustam R, Bernard H, Alfred R, Samejima H, Duckworth JW, Breitenmoser-Wuersten C, Belant JL, Hofer H, Wilting A. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and Distributions 19: 1366–1379.

Kress WJ, Garcia-Robledo C, Soares JVB, Jacobs D, Wilson K, Lopez IC, Belhumeur PN. 2018. Citizen Science and Climate Change: Mapping the Range Expansions of Native and Exotic Plants with the Mobile App Leafsnap. BioScience 68: 348–358.

Laaksonen M, Sajanti E, Sormunen JJ, Penttinen R, Hänninen J, Ruohomäki K, Sääksjärvi I, Vesterinen EJ, Vuorinen I, Hytönen J, Klemola T. 2017. Crowdsourcing-based nationwide tick collection reveals the distribution of Ixodes ricinus and I. persulcatus and associated pathogens in Finland. Emerging Microbes & Infections 6: 1–7.

Leibovici D, Williams J, Rosser J, Hodges C, Chapman C, Higgins C, Jackson M. 2017. Earth Observation for Citizen Science Validation, or Citizen Science for Earth Observation Validation? The Role of Quality Assurance of Volunteered Observations. Data 2: 35.

Leidenberger S, Boström S, Wayland M. 2020. Host records and geographical distribution of Corynosoma magdaleni, C. semerme and C. strumosum (Acanthocephala: Polymorphidae). Biodiversity Data Journal 8: e50500.

Lenanton RCJ, Dowling CE, Smith KA, Fairclough DV, Jackson G. 2017. Potential influence of a marine heatwave on range extensions of tropical fishes in the eastern Indian Ocean—Invaluable contributions from amateur observers. Regional Studies in Marine Science 13: 19–31.

Leung B, Hudgins EJ, Potapova A, Ruiz-Jaen MC. 2019. A new baseline for countrywide α-diversity and species distributions: illustration using >6,000 plant species in Panama. Ecological Applications 29: e01866.

Li E, Parker SS, Pauly GB, Randall JM, Brown BV, Cohen BS. 2019. An Urban Biodiversity Assessment Framework That Combines an Urban Habitat Classification Scheme and Citizen Science Data. Frontiers in Ecology and Evolution 7: 277.

Linhoss AC, Camacho R, Ashby S. 2016. Oyster Habitat Suitability in the Northern Gulf of Mexico. Journal of Shellfish Research 35: 841–849.

Longcore T, Noujdina N, Dixon PJ. 2018. Landscape Modeling of the Potential Natural Vegetation of Santa Catalina Island, California. Western North American Naturalist 78: 617.

Ma B, Sun J. 2018. Predicting the distribution of Stipa purpurea across the Tibetan Plateau via the MaxEnt model. BMC Ecology 18: 10.

MacLean MG, Congalton RG. 2015. A comparison of landscape fragmentation analysis programs for identifying possible invasive plant species locations in forest edge. Landscape Ecology 30: 1241–1256.

Mair L, Harrison PJ, Räty M, Bärring L, Strandberg G, Snäll T. 2017. Forest management could counteract distribution retractions forced by climate change. Ecological Applications 27: 1485–1497.

Maistrello L, Dioli P, Bariselli M, Mazzoli GL, Giacalone-Forini I. 2016. Citizen science and early detection of invasive species: phenology of first occurrences of Halyomorpha halys in Southern Europe. Biological Invasions 18: 3109–3116.

Mäkinen J, Vanhatalo J. 2018. Hierarchical Bayesian model reveals the distributional shifts of Arctic marine mammals. Diversity and Distributions 24: 1381–1394.

Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases?: Species diversity and distribution in the era of Big Data. Global Ecology and Biogeography 24: 973–984.

Malek R, Tattoni C, Ciolli M, Corradini S, Andreis D, Ibrahim A, Mazzoni V, Eriksson A, Anfora G. 2018. Coupling Traditional Monitoring and Citizen Science to Disentangle the Invasion of Halyomorpha halys. ISPRS International Journal of Geo-Information 7: 171.

Mammola S, Cardoso P, Angyal D, Balázs G, Blick T, Brustel H, Carter J, Ćurčić S, Danflous S, Dányi L, Déjean S, Deltshev C, Elverici M, Fernández J, Gasparo F, Komnenov M, Komposch C, Kováč L, Kunt K, Mock A, Moldovan O, Naumova M, Pavlek M, Prieto C, Ribera C, Rozwałka R, Růžička V, Vargovitsh R, Zaenker S, Isaia M. 2019. Continental data on cave-dwelling spider communities across Europe (Arachnida: Araneae). Biodiversity Data Journal 7: e38492.

Manenti R, Mori E, Di Canio V, Mercurio S, Picone M, Caffi M, Brambilla M, Ficetola GF, Rubolini D. 2020. The good, the bad and the ugly of COVID-19 lockdown effects on wildlife conservation: Insights from the first European locked down country. Biological Conservation 249: 108728.

Mang T, Essl F, Moser D, Karrer G, Kleinbauer I, Dullinger S. 2017. Accounting for imperfect observation and estimating true species distributions in modelling biological invasions. Ecography 40: 1187–1197.

Marcer A, Pino J, Pons X, Brotons L. 2012. Modelling invasive alien species distributions from digital biodiversity atlases. Model upscaling as a means of reconciling data at different scales. Diversity and Distributions 18: 1177–1189.

Marchante H, Morais MC, Gamela A, Marchante E. 2017. Using a WebMapping Platform to Engage Volunteers to Collect Data on Invasive Plants Distribution. Transactions in GIS 21: 238–252.

Marra S, de Lucia GA, Camedda A, Espinosa F, Coppa S. 2016. New records of the distribution and conservation status of the endangered limpet Patella ferruginea in Sardinia (Italy, W Mediterranean): Status and distribution of P. ferruginea in central-western Sardinia. Aquatic Conservation: Marine and Freshwater Ecosystems 26: 607–612.

Maslo B, Lockwood JL, Leu K. 2015. Land ownership patterns associated with declining forest birds: targeting the right policy and management for the right birds. Environmental Conservation 42: 216–226.

Mason SC, Hill JK, Thomas CD, Powney GD, Fox R, Brereton T, Oliver TH. 2018. Population variability in species can be deduced from opportunistic citizen science records: a case study using British butterflies. Insect Conservation and Diversity 11: 131–142.

da Mata RA, Tidon R, de Oliveira G, Vilela B, Diniz-Filho JAF, Rangel TF, Terribile LC. 2017. Stacked species distribution and macroecological models provide incongruent predictions of species richness for Drosophilidae in the Brazilian savanna. Insect Conservation and Diversity 10: 415–424.

McLeod SR, Pople AR. 2010. Modelling the distribution and relative abundance of feral camels in the Northern Territory using count data. The Rangeland Journal 32: 21.

van der Meer E. 2018. Carnivore conservation under land use change: the status of Zimbabwe's cheetah population after land reform. Biodiversity and Conservation 27: 647–663.

Metcalfe-Smith JL, Staton SK, Mackie GL, Lane NM. 1998. Selection of candidate species

of freshwater mussels (Bivalvia: Unionidae) to be considered for national status designation by COSEWIC. Canadian Field Naturalist 112: 425-440.

Mononen L, Auvinen A-P, Packalen P, Virkkala R, Valbuena R, Bohlin I, Valkama J, Vihervaara P. 2018. Usability of citizen science observations together with airborne laser scanning data in determining the habitat preferences of forest birds. Forest Ecology and Management 430: 498–508.

Monsarrat S, Novellie P, Rushworth I, Kerley G. Shifted distribution baselines: neglecting long-term biodiversity records risks overlooking potentially suitable habitat for conservation management. 374: 20190215.

Morán-Ordóñez A, Lahoz-Monfort JJ, Elith J, Wintle BA. 2017. Evaluating 318 continental-scale species distribution models over a 60-year prediction horizon: what factors influence the reliability of predictions?: Temporal transferability of species distribution model predictions. Global Ecology and Biogeography 26: 371–384.

Moro S, Jona-Lasinio G, Block B, Micheli F, De Leo G, Serena F, Bottaro M, Scacco U, Ferretti F. 2020. Abundance and distribution of the white shark in

the Mediterranean Sea. Fish and Fisheries 21: 338–349.
Morzaria-Luna HN, Cruz-Piñón G, Brusca RC, López-Ortiz AM, Moreno-Báez M, Reyes-Bonilla H, Turk-Boyer P. 2018. Biodiversity hotspots are not congruent with conservation areas in the Gulf of California. Biodiversity and Conservation 27: 3819–3842.
Mossman HL, Panter CJ, Dolman PM. 2015. Modelling biodiversity distribution in agricultural landscapes to support ecological network planning. Landscape and Urban Planning 141: 59–67.
Mostafavi H, Pletterbauer F, Coad BW, Mahini AS, Schinegger R, Unfer G, Trautwein C, Schmutz S. 2014. Predicting presence and absence of trout (Salmo trutta) in Iran. Limnologica 46: 1–8.
Moua Y, Roux E, Girod R, Dusfour I, de Thoisy B, Seyler F, Briolant S. 2016. Distribution of the Habitat Suitability of the Main Malaria Vector in French Guiana Using Maximum Entropy Modeling. Journal of Medical Entomology 54: 606-621.
Moulin N. 2020. When Citizen Science highlights alien invasive species in France: the case of Indochina mantis, Hierodula patellifera (Insecta, Mantodea, Mantidae). Biodiversity Data Journal 8: e46989.
Mueller MA, Drake D, Allen ML. 2019. Using citizen science to inform urban canid management. Landscape and Urban Planning 189: 362–371.
Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, Anderson RP. 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. Methods in Ecology and Evolution 5: 1198–1205.
Mweya CN, Kimera SI, Stanley G, Misinzo G, Mboera LEG. 2016. Climate Change Influences Potential Distribution of Infected Aedes aegypti Co-Occurrence with Dengue Epidemics Risk Areas in Tanzania. PLOS ONE 11: e0162649.
Myers AT, Gibbs JP. 2013. Landscape-level Factors Influencing Bog Turtle Persistence and Distribution in Southeastern New York State. Journal of Fish and Wildlife Management 4: 255–266.
Nimbs MJ, Willan RC, Smith SDA. 2017. Is Port Stephens, eastern Australia, a global hotspot for biodiversity of Aplysiidae (Gastropoda: Heterobranchia)? Molluscan Research 37: 45–65.
Norris D. 2014. Model Thresholds are More Important than Presence Location Type: Understanding the Distribution of Lowland tapir (Tapirus Terrestris) in a Continuous Atlantic Forest of Southeast Brazil. Tropical Conservation Science 7: 529–547.
Nottingham AC, Thompson JA, Wood F, Edwards PJ, Strager MP. 2019. Mapping pedomemory of spodic morphology using a species distribution model. Geoderma 352: 330–341.
Obidziński A, Pabjanek P, Medrzycki P. 2013. Determinants of badger Meles meles sett location in Białowieża Primeval Forest, northeastern Poland. Wildlife Biology 19: 48–68.

Obsomer V, Defourny P, Coosemans M. 2012. Predicted Distribution of Major Malaria Vectors Belonging to the Anopheles dirus Complex in Asia: Ecological Niche and Environmental Influences. PLoS ONE 7: e50475.

Ofori BY, Stow AJ, Baumgartner JB, Beaumont LJ. 2017. Combining dispersal, landscape connectivity and habitat suitability to assess climate-induced changes in the distribution of Cunningham's skink, Egernia cunninghami. PLOS ONE 12: e0184193.

Olson J, Wood J, Osborne R, Barrett-Lennard L, Larson S. 2018. Sightings of southern resident killer whales in the Salish Sea 1976-2014: the importance of a long-term opportunistic dataset. Endangered Species Research 37: 105–118.

Olsson O, Rogers DJ. 2009. Predicting the distribution of a suitable habitat for the white stork in Southern Sweden: identifying priority areas for reintroduction and habitat restoration. Animal Conservation 12: 62–70.

Osawa T, Watanabe K, Ikeda H, Yamamoto S. 2014. New approach for evaluating habitat stability using scarce records for both historical and contemporary specimens: a case study using Carabidae specimen records: Evaluating habitat with limited records. Entomological Science 17: 425–431.

Ostrowski M-F, Prosperi J-M, David J. 2016. Potential Implications of Climate Change on Aegilops Species Distribution: Sympatry of These Crop Wild Relatives with the Major European Crop Triticum aestivum and Conservation Issues. PLOS ONE 11: e0153974.

Ottaviani D, Lasinio GJ, Boitani L. 2004. Two statistical methods to validate habitat suitability models using presence-only data. Ecological Modelling 179: 417–443.

Oviedo L, Fernández M, Herra-Miranda D, Pacheco-Polanco JD, Hernández-Camacho CJ, Aurioles-Gamboa D. 2018. Habitat partitioning mediates the coexistence of sympatric dolphins in a tropical fjord-like embayment. Journal of Mammalogy 99: 554–564.

Pace DS, Giacomini G, Campana I, Paraboschi M, Pellegrino G, Silvestri M, Alessi J, Angeletti D, Cafaro V, Pavan G, Ardizzone G, Arcangeli A. 2019. An integrated approach for cetacean knowledge and conservation in the central Mediterranean Sea using research and social media data sources. Aquatic Conservation: Marine and Freshwater Ecosystems 29: 1302–1323.

Pantaleoni R. 2019. Going overseas: from island to continent colonization in the Mediterranean snakefly Fibla maclachlani (Albarda, 1891). BioInvasions Records 8: 442–451.

Pauly K, Jupp BP, De Clerck O. 2011. Modelling the distribution and ecology of Trichosolen blooms on coral reefs worldwide. Marine Biology 158: 2239–2246.

Pearson RG, Thuiller W, Araújo MB, Martinez-Meyer E, Brotons L, McClean C, Miles L, Segurado P, Dawson TP, Lees DC. 2006. Model-based uncertainty in species range prediction. Journal of Biogeography 33: 1704–1711.

Penado A, Rebelo H, Goulson D. 2016. Spatial distribution modelling reveals climatically suitable areas for bumblebees in undersampled parts of the Iberian

Peninsula. Insect Conservation and Diversity 9: 391–401.
Pereira AJ, Francisco A, Porto M. 2016. Flora-On: Occurrence data of the vascular flora of mainland Portugal. PhytoKeys 69: 105–119.
Périquet S, Roxburgh L, le Roux A, Collinson WJ. 2018. Testing the Value of Citizen Science for Roadkill Studies: A Case Study from South Africa. Frontiers in Ecology and Evolution 6: 15.
Pertierra LR, Bartlett JC, Duffy GA, Vega GC, Hughes KA, Hayward SAL, Convey P, Olalla-Tarraga MA, Aragón P. 2020. Combining correlative and mechanistic niche models with human activity data to elucidate the invasive potential of a sub-Antarctic insect. Journal of Biogeography 47: 658–673.
Pettorelli N, Hilborn A, Broekhuis F, Durant SM. 2009. Exploring habitat use by cheetahs using ecological niche factor analysis. Journal of Zoology 277: 141–148.
Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19: 181–197.
Phipps WL, Diekmann M, MacTavish LM, Mendelsohn JM, Naidoo V, Wolter K, Yarnell RW. 2017. Due South: A first assessment of the potential impacts of climate change on Cape vulture occurrence. Biological Conservation 210: 16–25.
Platenberg RJ, Harvey DS. Endangered species and land use conflicts: a case study of the Virgin Islands boa (Epicrates granti). Herpetological Conservation and Biology 5: 548-554.
Polgar CA, Primack RB, Williams EH, Stichter S, Hitchcock C. 2013. Climate effects on the flight period of Lycaenid butterflies in Massachusetts. Biological Conservation 160: 25–31.
Porter WT, Motyka PJ, Wachara J, Barrand ZA, Hmood Z, McLaughlin M, Pemberton K, Nieto NC. 2019. Citizen science informs human-tick exposure in the Northeastern United States. International Journal of Health Geographics 18: 9.
Puan CL, Yong DL, Azhar B, Phua MH, Lim KC. 2015. Ecological correlations of nocturnal bird assemblages in Malaysian Borneo. Forktail 31: 82-86.
Qin Z, Zhang JE, Jiang YP, Wang RL, Wu RS. 2020. Predicting the potential distribution of Pseudomonas syringae pv. actinidiae in China using ensemble models. Plant Pathology 69: 120–131.
Quevedo M, Bañuelos MJ, Obeso JR. 2006. The decline of Cantabrian capercaillie: How much does habitat configuration matter? Biological Conservation 127: 190–200.
Rapacciuolo G, Ball-Damerow JE, Zeilinger AR, Resh VH. 2017. Detecting long-term occupancy changes in Californian odonates from natural history and citizen science records. Biodiversity and Conservation 26: 2933–2949.
Ready J, Kaschner K, South AB, Eastwood PD, Rees T, Rius J, Agbayani E, Kullander S, Froese R. 2010. Predicting the distributions of marine organisms

at the global scale. Ecological Modelling 221: 467–478.
Rebelo H, Jones G. 2010. Ground validation of presence-only modelling with rare species: a case study on barbastelles Barbastella barbastellus (Chiroptera: Vespertilionidae). Journal of Applied Ecology 47: 410–420.
Rebelo H, Tarroso P, Jones G. 2010. Predicted impact of climate change on European bats in relation to their biogeographic patterns. Global Change Biology 16: 561–576.
Redhead JW, Fox R, Brereton T, Oliver TH. 2016. Assessing species' habitat associations from occurrence records, standardised monitoring data and expert opinion: A test with British butterflies. Ecological Indicators 62: 271–278.
Reese GC, Skagen SK. 2017. Modeling nonbreeding distributions of shorebirds and waterfowl in response to climate change. Ecology and Evolution 7: 1497–1513.
Ribas LG dos S, de Cássia-Silva C, Petsch DK, Silveira MJ, Lima-Ribeiro MS. 2018. The potential invasiveness of an aquatic macrophyte reflects founder effects from native niche. Biological Invasions 20: 3347–3355.
Ringvold (HR) H, Hatlevik (AH) A, Hevrøy (JH) J, Hughes (MH) M, Aukan (NA) N. 2020. Encounters with the rare genus Helicosalpa (Chordata, Thaliacea, Salpida), using citizen science data. Marine Biology Research 16: 369–379.
Robinson JL, Fordyce JA, Parker CR. 2016. Conservation of aquatic insect species across a protected area network: null model reveals shortfalls of biogeographical knowledge. Journal of Insect Conservation 20: 565–581.
Rockhill AP, Sollman R, Powell RA, DePerno CS. 2016. A Comparison of Survey Techniques for Medium- to Large-Sized Mammals in Forested Wetlands. Southeastern Naturalist 15: 175–187.
Rodríguez-Soto C, Monroy-Vilchis O, Maiorano L, Boitani L, Faller JC, Briones MÁ, Núñez R, Rosas-Rosas O, Ceballos G, Falcucci A. 2011. Predicting potential distribution of the jaguar (Panthera onca) in Mexico: identification of priority areas for conservation: Jaguar potential distribution in Mexico. Diversity and Distributions 17: 350–361.
Routson KJ, Volk GM, Richards CM, Smith SE, Nabhan GP, Wyllie de Echeverria V. 2012. Genetic Variation and Distribution of Pacific Crabapple. Journal of the American Society for Horticultural Science 137: 325–332.
Rumi A, Vogler RE, Beltramino AA. 2017. The South-American distribution and southernmost record of Biomphalaria peregrina —a potential intermediate host of schistosomiasis. PeerJ 5: e3401.
Saito M, Momose H, Mihira T, Uematsu S. 2012. Predicting the risk of wild boar damage to rice paddies using presence-only data in Chiba Prefecture, Japan. International Journal of Pest Management 58: 65–71.
Sallam MF, Al Ahmed AM, Abdel-Dayem MS, Abdullah MAR. 2013. Ecological Niche Modeling and Land Cover Risk Areas for Rift Valley Fever Vector, Culex tritaeniorhynchus Giles in Jazan, Saudi Arabia. PLoS ONE 8: e65786.
Santori C, Spencer R-J, Van Dyke JU, Thompson MB. 2018. Road mortality of the eastern long-necked turtle (Chelodina longicollis) along the Murray River,

Australia: an assessment using citizen science. Australian Journal of Zoology 66: 41.

Sardà-Palomera F, Vieites DR. 2011. Modelling Species' Climatic Distributions Under Habitat Constraints: A Case Study with Coturnix coturnix. Annales Zoologici Fennici 48: 147–160.

Scarnati L, Attorre F, Farcomeni A, Francesconi F, Sanctis M. 2009. Modelling the spatial distribution of tree species with fragmented populations from abundance data. Community Ecology 10: 215–224.

Scholte RGC, Carvalho OS, Malone JB, Utzinger J, Vounatsou P. 2012. Spatial distribution of Biomphalaria spp., the intermediate host snails of Schistosoma mansoni, in Brazil. Geospatial health 6: 95.

Schubert SC, Manica LT, Guaraldo ADC. 2019. Revealing the potential of a huge citizen-science platform to study bird migration. Emu - Austral Ornithology 119: 364–373.

Senay SD, Worner SP, Ikeda T. 2013. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. PLoS ONE 8: e71218.

Shabani F, Kumar L. 2015. Should species distribution models use only native or exotic records of existence or both? Ecological Informatics 29: 57–65.

Sharifi M, Karami P, Akmali V, Afroosheh M, Vaissi S. Modeling Geographic Distribution for the Endangered Yellow Spotted Mountain Newt, Neurergus microspilotus (Amphibia: Salamandridae) in Iran and Iraq. Herpetological Conservation and Biology 12: 488–497.

Sharma S, Arunachalam K, Bhavsar D, Kala R. 2018. Modeling habitat suitability of Perilla frutescens with MaxEnt in Uttarakhand—A conservation approach. Journal of Applied Research on Medicinal and Aromatic Plants 10: 99–105.

Shatz AJ, Rogan J, Sangermano F, Ogneva-Himmelberger Y, Chen H. 2013. Characterizing the potential distribution of the invasive Asian longhorned beetle (Anoplophora glabripennis) in Worcester County, Massachusetts. Applied Geography 45: 259–268.

Shirakihara M. 2012. Long-Distance Movements of Indo-Pacific Bottlenose Dolphins (Tursiops aduncus) and Habitat Preference of Two Species of Bottlenose Dolphins in Eastern Kyushu, Japan. Aquatic Mammals 38: 145–152.

Shumba T, Montgomery RA, Rasmussen GSA, Macdonald DW. 2018. African Wild Dog Habitat Use Modelling Using Telemetry Data and Citizen Scientist Sightings: Are the Results Comparable? African Journal of Wildlife Research 48: 013002.

Sindato C, Stevens KB, Karimuribo ED, Mboera LEG, Paweska JT, Pfeiffer DU. 2016. Spatial Heterogeneity of Habitat Suitability for Rift Valley Fever Occurrence in Tanzania: An Ecological Niche Modelling Approach. PLOS Neglected Tropical Diseases 10: e0005002.

Skejo J, Deranja M, Adžić K. 2020. Pygmy Hunchback of New Caledonia: Notredamia dora gen. n. et sp. n. – A New Cladonotin (Caelifera: Tetrigidae) Genus and Species from Oceania. Entomological News 129: 170.

Soley-Guardia M, Gutiérrez EE, Thomas DM, Ochoa-G J, Aguilera M, Anderson RP. 2016. Are we overestimating the niche? Removing marginal localities helps ecological niche models detect environmental barriers. Ecology and Evolution 6: 1267–1279.

Soroye P, Ahmed N, Kerr JT. 2018. Opportunistic citizen science data transform understanding of species distributions, phenology, and diversity gradients for global change research. Global Change Biology 24: 5281–5291.

Sousa-Baena MS, Garcia LC, Townsend Peterson A. 2014. Knowledge behind conservation status decisions: Data basis for "Data Deficient" Brazilian plant species. Biological Conservation 173: 80–89.

Speed JDM, Bendiksby M, Finstad AG, Hassel K, Kolstad AL, Prestø T. 2018. Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. PLOS ONE 13: e0196417.

Stas M, Aerts R, Hendrickx M, Dendoncker N, Dujardin S, Linard C, Nawrot T, Van Nieuwenhuyse A, Aerts J-M, Van Orshoven J, Somers B. 2020. An evaluation of species distribution models to estimate tree diversity at genus level in a heterogeneous urban-rural landscape. Landscape and Urban Planning 198: 103770.

Stafford R, Hart AG, Goodenough AE. 2013. A visual method to identify significant latitudinal changes in species' distributions. Ecological Informatics 15: 74–84.

Stefanescu C, Páramo F, Åkesson S, Alarcón M, Ávila A, Brereton T, Carnicer J, Cassar LF, Fox R, Heliölä J, Hill JK, Hirneisen N, Kjellén N, Kühn E, Kuussaari M, Leskinen M, Liechti F, Musche M, Regan EC, Reynolds DR, Roy DB, Ryrholm N, Schmaljohann H, Settele J, Thomas CD, van Swaay C, Chapman JW. 2013. Multi-generational long-distance migration of insects: studying the painted lady butterfly in the Western Palaearctic. Ecography 36: 474–486.

Stone AI, Lima EM, Aguiar GFS, Camargo CC, Flores TA, Kelt DA, Marques-Aguiar SA, Queiroz JAL, Ramos RM, Silva Júnior JS. 2009. Non-volant mammalian diversity in fragments in extreme eastern Amazonia. Biodiversity and Conservation 18: 1685–1694.

Strebel N, Kéry M, Schaub M, Schmid H. 2014. Studying phenology by flexible modelling of seasonal detectability peaks. Methods in Ecology and Evolution 5: 483–490.

Supp SR, La Sorte FA, Cormier TA, Lim MCW, Powers DR, Wethington SM, Goetz S, Graham CH. 2015. Citizen-science data provides new insight into annual and seasonal variation in migration patterns. Ecosphere 6: 15.

van Strien AJ, Boomsluiter M, Noordeloos ME, Verweij RJT, Kuyper TW. 2018. Woodland ectomycorrhizal fungi benefit from large-scale reduction in nitrogen deposition in the Netherlands. Journal of Applied Ecology 55: 290–298.

van Strien AJ, van Swaay CAM, Termaat T. 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. Journal of Applied Ecology 50: 1450–1458.

Suazo CG, Yates O, Azócar J, Díaz P, González-But JC, Cabezas LA. 2017. Emerging platforms to monitor the occurrence and threats to critically endangered seabirds: The waved albatross in Chile and the Southeast Pacific. Revista de biología marina y oceanografía 52: 245–254.

Sun Y, Li L, Li L, Zou J, Liu J. 2015. Distributional dynamics and interspecific gene flow in Picea likiangensis and P. wilsonii triggered by climate change on the Qinghai-Tibet Plateau. Journal of Biogeography 42: 475–484.

Syfert MM, Smith MJ, Coomes DA. 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. PLoS ONE 8: e55158.

Szabo JK, Vesk PA, Baxter PWJ, Possingham HP. 2010. Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. Ecological Applications 20: 2157–2169.

Szabo ND, Colla SR, Wagner DL, Gall LF, Kerr JT. 2012. Do pathogen spillover, pesticide use, or habitat loss explain recent North American bumblebee declines?: Causes of bumblebee declines. Conservation Letters 5: 232–239.

Tanner AM, Tanner EP, Papeş M, Fuhlendorf SD, Elmore RD, Davis CA. 2020. Using aerial surveys and citizen science to create species distribution models for an imperiled grouse. Biodiversity and Conservation 29: 967–986.

Tantipisanuh N, Gale GA. 2018. Identification of biodiversity hotspot in national level – Importance of unpublished data. Global Ecology and Conservation 13: e00377.

Termaat T, van Strien AJ, van Grunsven RHA, De Knijf G, Bjelke U, Burbach K, Conze K, Goffart P, Hepper D, Kalkman VJ, Motte G, Prins MD, Prunier F, Sparrow D, van den Top GG, Vanappelghem C, Winterholler M, WallisDeVries MF. 2019. Distribution trends of European dragonflies under climate change. Diversity and Distributions 25: 936–950.

Thompson PM, Brookes KL, Cordes LS. 2015. Integrating passive acoustic and visual data to model spatial patterns of occurrence in coastal dolphins. ICES Journal of Marine Science 72: 651–660.

Tiffin HS, Peper ST, Wilson-Fallon AN, Haydett KM, Cao G, Presley SM. 2019. The Influence of New Surveillance Data on Predictive Species Distribution Modeling of Aedes aegypti and Aedes albopictus in the United States. Insects 10: 400.

Tingley R, Thompson MB, Hartley S, Chapple DG. 2016. Patterns of niche filling and expansion across the invaded ranges of an Australian lizard. Ecography 39: 270–280.

Tovaranonte J, Blach-Overgaard A, Pongsattayapipat R, Svenning J-C, Barfod AS. 2015. Distribution and diversity of palms in a tropical biodiversity hotspot (Thailand) assessed by species distribution modeling. Nordic Journal of Botany 33: 214–224.

Townsend Peterson A, Ortega Huerta MA. 2008. Modelado de nichos ecológicos y predicción de distribuciones geográficas: comparación de seis métodos. Revista

Mexicana de Biodiversidad 79: 205-216.
Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific Reports 7: 9132.
Václavík T, Meentemeyer RK. 2009. Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? Ecological Modelling 220: 3248–3258.
Vale MM, Tourinho L, Lorini ML, Rajão H, Figueiredo MSL. 2018. Endemic birds of the Atlantic Forest: traits, conservation status, and patterns of biodiversity. Journal of Field Ornithology 89: 193–206.
Vásquez-Ordóñez AA, Hazzi NA, Escobar-Prieto D, Paz-Jojoa D, Parsa S. 2015. A geographic distribution database of the Neotropical cassava whitefly complex (Hemiptera, Aleyrodidae) and their associated parasitoids and hyperparasitoids (Hymenoptera). ZooKeys 545: 75–87.
Vergara J, Acosta LE. 2015. More on the Mesopotamian-Yungas disjunction in subtropical and temperate Argentina: Bioclimatic distribution models of the harvestman Discocyrtus dilatatus (Opiliones: Gonyleptidae). Zoologia (Curitiba) 32: 445–456.
Verlaque M, Breton G. 2019. Biological invasion: Long term monitoring of the macroalgal flora of a major European harbor complex. Marine Pollution Bulletin 143: 228–241.
Vihervaara P, Mononen L, Auvinen A-P, Virkkala R, Lü Y, Pippuri I, Packalen P, Valbuena R, Valkama J. 2015. How to integrate remotely sensed data and biodiversity for ecosystem assessments at landscape scale. Landscape Ecology 30: 501–516.
Villordon A, Njuguna W, Gichuki S, Ndolo P, Kulembeka H, Jeremiah SC, LaBonte D, Yada B, Tukamuhabwa P, Mwanga ROM. 2006. Using GIS-Based Tools and Distribution Modeling to Determine Sweetpotato Germplasm Exploration and Documentation Priorities in Sub-Saharan Africa. HortScience 41: 1377–1381.
Vray S, Rollin O, Rasmont P, Dufrêne M, Michez D, Dendoncker N. 2019. A century of local changes in bumblebee communities and landscape composition in Belgium. Journal of Insect Conservation 23: 489–501.
Wan X, Jiang G, Yan C, He F, Wen R, Gu J, Li X, Ma J, Stenseth NChr, Zhang Z. 2019. Historical records reveal the distinctive associations of human disturbance and extreme climate change with local extinction of mammals. Proceedings of the National Academy of Sciences 116: 19001–19008.
Wang Y, Casajus N, Buddle C, Berteaux D, Larrivée M. 2018. Predicting the distribution of poorly-documented species, Northern black widow (Latrodectus variolus) and Black purse-web spider (Sphodros niger), using museum specimens and citizen science data. PLOS ONE 13: e0201094.
Wang Y-H, Yang K-C, Bridgman CL, Lin L-K. 2008. Habitat suitability modelling to correlate gene flow with landscape connectivity. Landscape Ecology s10980-008-9262–3.

Warton DI, Renner IW, Ramp D. 2013. Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. PLoS ONE 8: e79168.

Washitani I, Nagai M, Yasukawa M, Kitsuregawa M. 2020. Testing a butterfly commonness hypothesis with data assembled by a citizen science program "Tokyo Butterfly Monitoring". Ecological Research 35: 1087–1094.

Watts SM, McCarthy TM, Namgail T. 2019. Modelling potential habitat for snow leopards (Panthera uncia) in Ladakh, India. PLOS ONE 14: e0211509.

Webb MH, Terauds A, Tulloch A, Bell P, Stojanovic D, Heinsohn R. 2017. The importance of incorporating functional habitats into conservation planning for highly mobile species in dynamic systems: Habitat Function in Dynamic Systems. Conservation Biology 31: 1018–1028.

Webster RP, de Tonnancour P, Sweeney JD, Webster VL, Kostanowicz CA, Hughes C, Anderson RS, Klymko J, Chantal C, Vigneault R. 2020. New Coleoptera records from eastern Canada, with additions to the fauna of Manitoba, British Columbia, and Yukon Territory. ZooKeys 946: 53–112.

Wei B, Wang R, Hou K, Wang X, Wu W. 2018. Predicting the current and future cultivation regions of Carthamus tinctorius L. using MaxEnt model under climate change in China. Global Ecology and Conservation 16: e00477.

Wepfer PH, Guénard B, Economo EP. 2016. Influences of climate and historical land connectivity on ant beta diversity in East Asia. Journal of Biogeography 43: 2311–2321.

Westwood R, Westwood AR, Hooshmandi M, Pearson K, LaFrance K, Murray C. 2020. A field-validated species distribution model to support management of the critically endangered Poweshiek skipperling (Oarisma poweshiek) butterfly in Canada. Conservation Science and Practice 2.

Whiting SD, Macrae I, Thorn R, Murray W, Whiting AU. 2014. Sea turtles of the Cocos (Keeling) Islands, Indian Ocean. The Raffles Bulletin of Zoology 30: 168-183.

Williams HF, Bartholomew DC, Amakobe B, Githiru M. 2018. Environmental factors affecting the distribution of African elephants in the Kasigau wildlife corridor, SE Kenya. African Journal of Ecology 56: 244–253.

Wilting A, Hearn AJ, Eaton J, Belant JL, Kramer-Schadt S. 2016. Predicted distribution of the Bornean ferret badger Melogale everetti (Mammalia: Carnivora: Mustelidae) on Borneo. The Raffles Bulletin of Zoology 33: 55-60.

Witt M, Hardy T, Johnson L, McClellan C, Pikesley S, Ranger S, Richardson P, Solandt J, Speedie C, Williams R, Godley B. 2012. Basking sharks in the northeast Atlantic: spatio-temporal trends from sightings in UK waters. Marine Ecology Progress Series 459: 121–134.

Wonham MJ, Hart MW. 2018. El Niño Range Extensions of Pacific Sand Crab (Emerita analoga) in the Northeastern Pacific. Northwest Science 92: 53–60.

Xing S, Au TF, Dufour PC, Cheng W, Landry Yuan F, Jia F, Vu LV, Wang M, Bonebrake TC. 2019. Conservation of data deficient species under multiple threats: Lessons from an iconic tropical butterfly (Teinopalpus aureus).

Biological Conservation 234: 154–164.
Yu F, Wang T, Groen TA, Skidmore AK, Yang X, Ma K, Wu Z. 2019. Climate and land use changes will degrade the distribution of Rhododendrons in China. Science of The Total Environment 659: 515–528.
Yue S, Bonebrake TC, Gibson L. 2019. Informing snake roadkill mitigation strategies in Taiwan using citizen science: Snake Roadkill Mitigation Using Citizen Science. The Journal of Wildlife Management 83: 80–88.
Zaniewski AE, Lehmann A, Overton JM. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157: 261–280.
Zapponi L, Cini A, Bardiani M, Hardersen S, Maura M, Maurizi E, Redolfi De Zan L, Audisio P, Bologna MA, Carpaneto GM, Roversi PF, Sabbatini Peverieri G, Mason F, Campanaro A. 2017. Citizen science data as an efficient tool for mapping protected saproxylic beetles. Biological Conservation 208: 139–145.
Zarzo-Arias A, Penteriani V, Delgado M del M, Peón Torre P, García-González R, Mateo-Sánchez MC, Vázquez García P, Dalerum F. 2019. Identifying potential areas of expansion for the endangered brown bear (Ursus arctos) population in the Cantabrian Mountains (NW Spain). PLOS ONE 14: e0209972.
Zimmermann H, Von Wehrden H, Damascos MA, Bran D, Welk E, Renison D, Hensen I. 2011. Habitat invasion risk assessment based on Landsat 5 data, exemplified by the shrub Rosa rubiginosa in southern Argentina: HABITAT INVASION RISK ASSESSMENT. Austral Ecology 36: 870–880.

S8. R scripts

All scripts used in data management and analysis for our review are available at https://doi.org/10.17605/OSF.IO/JUEQC.

S9. References for supplementary materials

1. Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* **339.** https://doi.org/10.1136/bmj.b2535 (2009).
2. Aria, M. & Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics* **11,** 959–975. https://doi.org/10.1016/j.joi.2017.08.007 (2017).
3. Westgate, M. J. revtools: An R package to support article screening for evidence synthesis. *Research synthesis methods* **10,** 606–614. https://doi.org/10.1002/jrsm.1374 (2019).
4. GBIF. GBIF science review 2019. *Global Biodiversity Information Facility.* https://doi.org/10.15468/QXXG-7K93 (2019).

Paper II

SCIENTIFIC REPORTS, VOLUME 12, ARTICLE NUMBER: 7648 (2022), DOI:10.1038/S41598-022-11257-X

Maximizing citizen scientists' contribution to automated species recognition

Wouter Koch[1,2] Laurens Hogeweg[3,4] Erlend B. Nilsen[5]
Anders G. Finstad[1]

1. Department of Natural History, Norwegian University of Science and Technology, Trondheim, Norway; 2. Norwegian Biodiversity Information Centre, Trondheim, Norway; 3. Intel Benelux, High Tech Campus 83, 5656 AE Eindhoven, The Netherlands; 4. Naturalis Biodiversity Center, PO Box 9517, 2300 RA, Leiden, The Netherlands; 5. Faculty of Biosciences and Aquaculture, Nord University, Steinkjer, Norway

Published: 10 May 2022

Abstract

Technological advances and data availability have enabled artificial intelligence-driven tools that can increasingly successfully assist in identifying species from images. Especially within citizen science, an emerging source of information filling the knowledge gaps needed to solve the biodiversity crisis, such tools can allow participants to recognize and report more poorly known species. This can be an important tool in addressing the substantial taxonomic bias in biodiversity data, where broadly recognized, charismatic species are highly over-represented. Meanwhile, the recognition models are trained using the same biased data, so it is important to consider what additional images are needed to improve recognition models. In this study, we investigated how the amount of training data influenced the performance of species recognition models for various taxa. We utilized a large citizen science dataset collected in Norway, where images are added independently from identification. We demonstrate that while adding images of currently under-represented taxa will generally improve recognition models more, there are important deviations from this general pattern. Thus, a more focused prioritization of data collection beyond the basic paradigm that "more is better" is likely to significantly improve species recognition models and advance the representativeness of biodiversity data.

Keywords: Image recognition, Taxonomic bias, Value of Information, Citizen science

Introduction

Addressing the current crisis related to biodiversity loss necessarily involves addressing several fundamental knowledge gaps[1,2]. Currently there are vast spatial, temporal and especially taxonomic gaps and biases in global primary biodiversity data sets, limiting our understanding of the earth's biosphere[3–6]. Several observation methods based on image recognition, ranging from remotely operated vessels to camera traps and citizen science programs[7–9], hold great promise in solving some of the taxonomic biases currently experienced[10]. Citizen science (observations made by non-professional volunteers[11]) has emerged as a very large source of biodiversity data. It has the potential to fill gaps in our current knowledge about the occurrence of species in time and space[12–14]. Several citizen science programs, e.g. iNaturalist, eBird, iSpot[15] contribute data on vast scales and in amounts that cannot feasibly be acquired in any other way. Such programs come with the added benefit of educating and engaging the general public[16–18]. Some of the main concerns related to citizen science data are reliability of the taxon identifications reported[19,20], and the over-representation of more charismatic taxa such as birds and flowering plants[21–23]. Improving the quality of citizen science data is a vital step in addressing the knowledge gaps in our understanding of the earth's biosphere.

Image recognition models can help citizen scientists recognize more species and provide a quality control mechanism that helps to reduce the risk of species misidentification[10]. Their performance is however inherently linked to the quality of the data used to train them. By increasingly helping citizen scientists identify species from images[24–26], such tools help address the aforementioned issues in citizen science data, adding to the quality, quantity and taxonomic scope of observations. Image recognition models can warn the citizen scientist and validators of potential misclassifications. Output of citizen scientists is increased as automating parts of the reporting process makes reporting less time consuming. Image recognition models also allow citizen scientists to report more of what they encounter by enabling them to report taxa they could not have identified independently. The taxonomic scope of the citizen scientist is expanded when tools enable them to identify and report within taxa they would otherwise not be familiar with[27]. Observations accompanied by images can be used for training an image recognition model for use in the field. Generally one aims to keep training data as similar as possible to the intended classification task of the model[28]. By using images from citizen scientists when training a model intended for use within citizen science, one is more likely to capture the variability in the kind of images provided by citizen scientists. Images from other sources may be more standardized, depict close-ups of relevant features, and/or depict preserved and prepared specimens. Deep neural networks are designed to draw inferences from novel data by generalized patterns observed in training data[28], requiring substantial amounts of data. The Computer Vision model by iNaturalist, for example, only includes taxa for which

at least 100 images are available[29]. This criterion excludes 89% of the taxa with at least one image in the dataset used in this study, illustrating how heavily the training of models depends on the amounts of data citizen science provides. In this manner, citizen science and automated image recognition are increasingly interdependent. Image recognition models help citizen scientists collect data to expand our knowledge base, whilst training of the next generation of recognition models depends on the collection of more images.

While some species are readily recognized with limited experience, others require extensive experience with many specimens to obtain the necessary knowledge. The distinct colorations of butterflies may allow any interested observer with some experience to reliably identify the majority of species in certain areas, while a taxon like Diptera remains notoriously difficult even after years of study. Machine learning is no different from human learning in this respect; different amounts of training data are required depending on the distinctness of species' characteristics. Therefore, there can be substantial differences between taxa in the number of images required per species for the best achievable model performance. This can depend on factors like species' distinctiveness, the variation in appearance, the various angles and contexts in which photos are taken, and the extent to which a species' behavior is suited for high quality documentation.[27,30,31] As a result, the value of adding a new image to the training set is not equal across taxa, but varies both because the size of the existing training set is different, and the fact that some species are more distinct than others. Thus it is important to consider the informational value of adding images to the training data.

In this study, we use the Species Observation Service, a large Norwegian citizen science project, as an example to investigate the nature of the bias in citizen science image data, and how this relates to the value of data for image recognition models. One way to evaluate this is by using the concept of Value of Information (VoI); "*the increase in expected value that arises from making the best choice with the benefit of a piece of information compared to the best choice without the benefit of that same information*"[32]. Considering training data for image recognition models in the VoI framework allows us to identify the most effective prioritization for improving recognition models. This method allows for a more sophisticated approach to data collection than simply adding more data for all taxa, or even for taxa that are currently the most under-represented. First, we evaluate whether the biases generally found in all observation data, regardless of source, are the same within citizen science observations with images, or if there are different biases that need to be taken into account. Then we train multiple image recognition models for different taxa, with a gradually increasing number of images per species, allowing us to quantify and compare the effects of adding more training data between taxa. Using these changes in performance, we estimate the VoI of adding training data for each taxon, relative to the amount of images that are currently available. Finally, comparing this VoI to the amount of over- or under-representation of these taxa, we demonstrate that mobilizing images with a higher VoI provides an

alternative, data-driven and efficient approach compared to simply prioritizing images of the currently most under-represented taxa.

Results

Taxonomic bias in citizen science observations with or without images

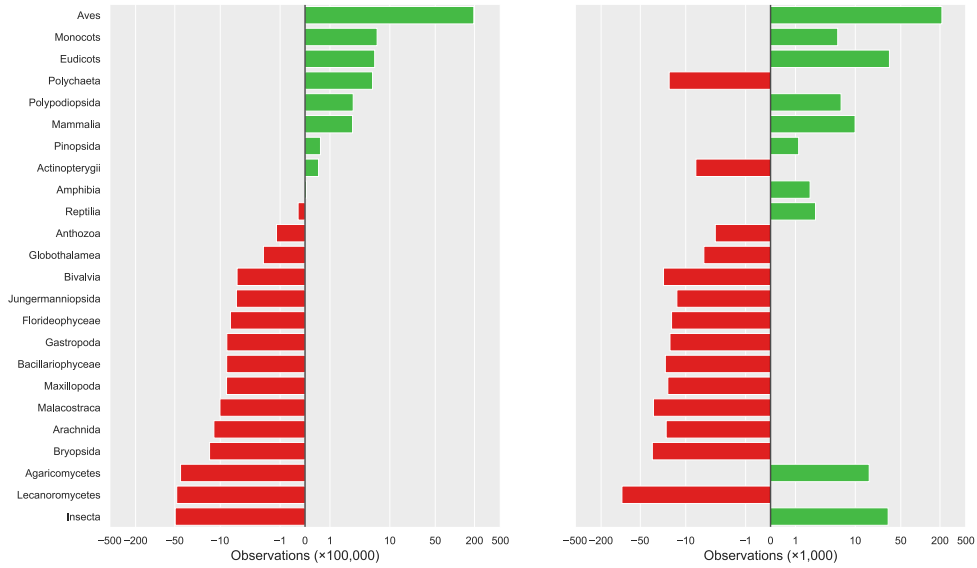
It has been well documented in a global context that particularly charismatic taxonomic classes have many times more reported observations per species than those that are considered less charismatic[5]. We find the same pattern when considering classes within the totality of GBIF mediated observations for Norway from all sources (figure 1a). When limiting this analysis to only observations with images that originate from the citizen science platform Species Observation Service[33], a different pattern emerges (figure 1b). Perhaps most eye-catchingly, Insecta are the most under-represented taxon in the totality of Norwegian observations, but the 3rd most over-represented when limiting the analysis to citizen science images. We performed a similar analysis for the 12 taxonomic orders used in the machine learning part of this study. This provides the biases in relative representation per species in the data available for training our recognition models.

Image recognition performance and the Value of Information

When training image recognition models, the amount of training data provided to the model determines how well the model is able to recognize species in the test images. For all orders, as models are provided with more images per species, their performance (as measured by the F_1 scores) increases. Comparing the performances for each order at the lowest and highest number of training images per species, as well as the gradual performance increase over intermediate numbers of training images, it is clear that the 12 orders have distinct performance curves (figure 2). From this it follows that the increase in performance at any given point on these curves - the Value of Information (VoI) of adding observations with images at that point - also differs between orders. Combined with substantially different amounts of currently available observations between orders, the estimated VoI of adding an observation with at least one image to those currently available for that order also varies widely (figure 3).

Combining Value of Information and taxonomic bias

After obtaining the per-species over- and under-representations, as well as the current expected VoI of additional observations with images, we can compare the two values for each order in the experiment. Plotting the taxonomic bias of the orders used in this experiment together with their estimates for their respective



(a) (b)

Figure 1: The per-species representation of observations in Norway per class, using all GBIF data (a) or only GBIF mediated citizen science data with images (b). The 0-line is where the values would be if the average number of observations per species in that class was equal to the average number of observations per species over all classes combined. Plotted here on an inverse hyperbolic sine-transformed scale, sorted by the per-species representation in subplot (a).

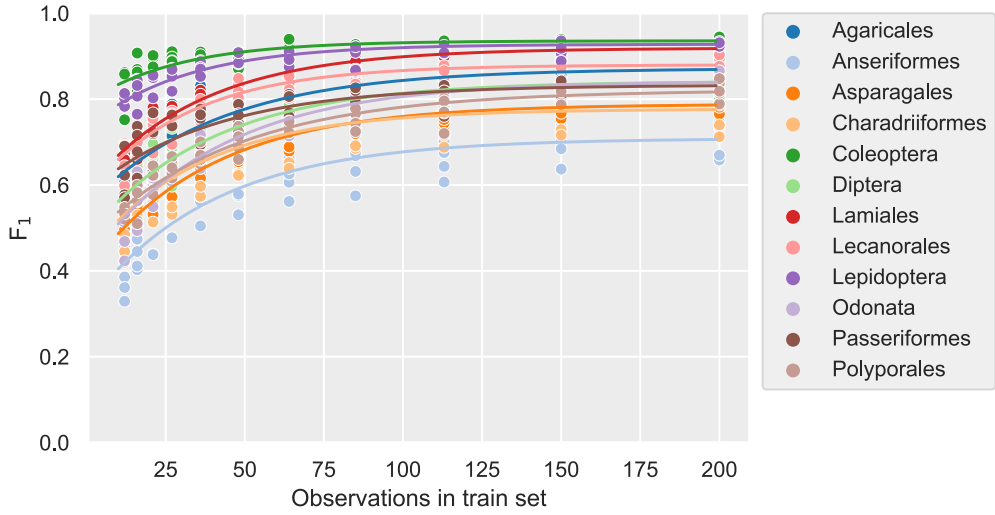


Figure 2: The performance (F_1 score) vs the training set size. Lines are the fitted Von Bertalanffy Growth Function-curves per order. See the Supplementary Information for an interpretation of such curves.

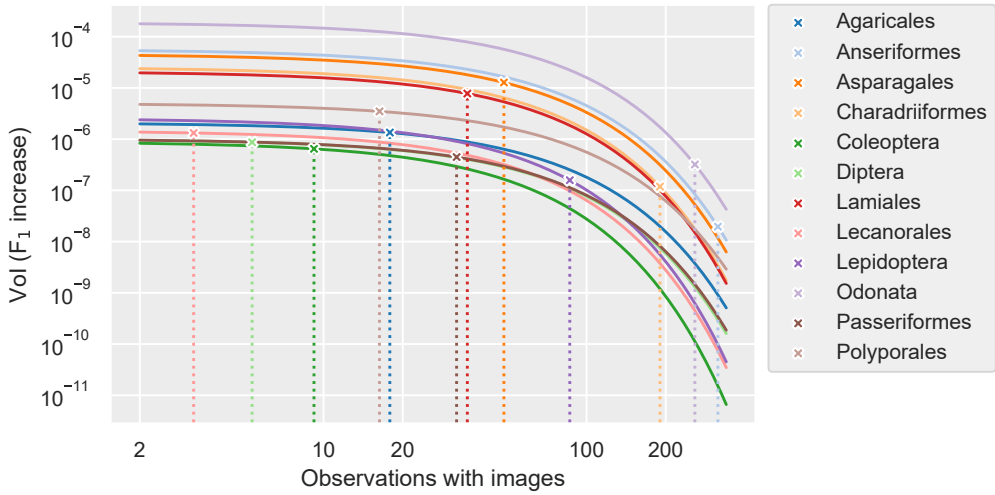


Figure 3: The VoI (F_1 increase) for each order as the result of adding a single observation with at least one image for a single species, versus the average number of observations with images available per species. Dotted lines mark the average number of observations with images per species currently available for the respective order, from which the current expected VoI (marked with x) is derived.

estimated VoI, it is clear that current under- or over-representation of the order is not the determining factor for the expected value of additional observations. While the VoI of under-represented orders is generally higher, differences between

orders in their learning curves cause some orders to have a higher or lower VoI than just their overall over- or under-representation would indicate (figure 4).

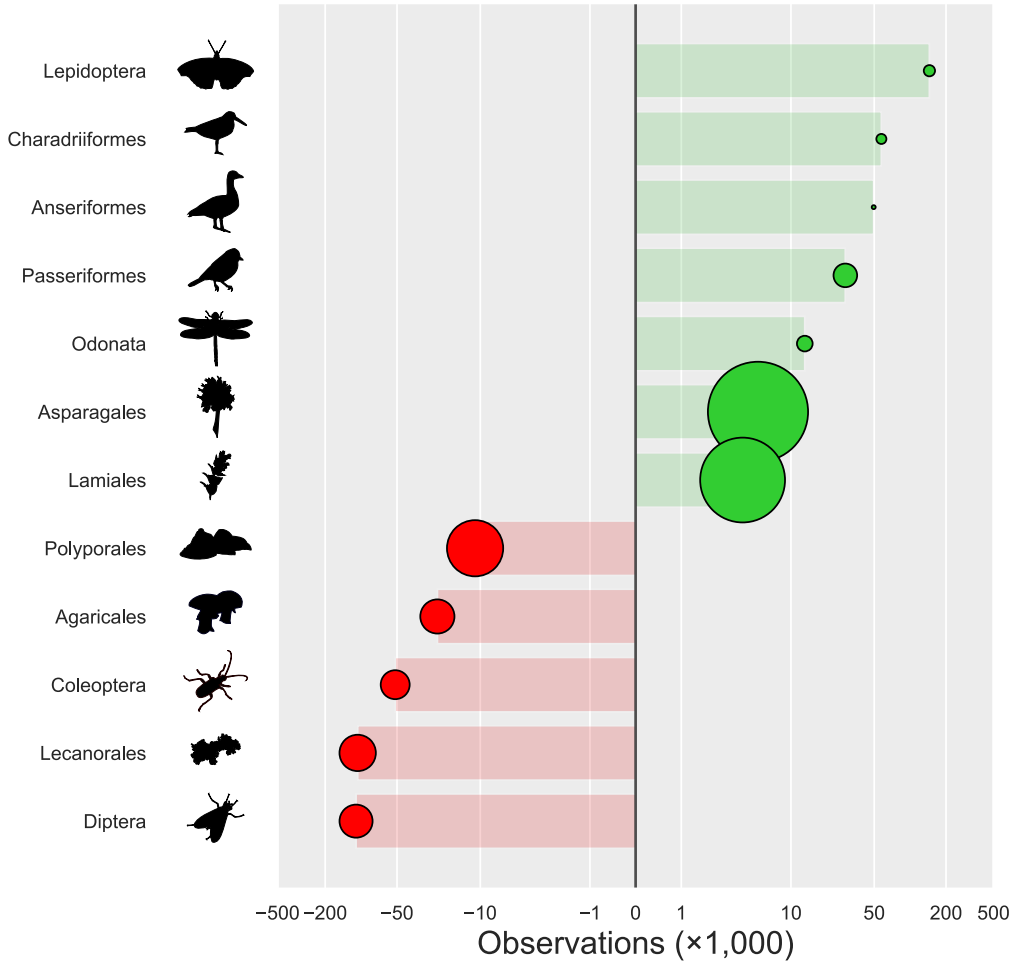


Figure 4: The relative per-species representation in Norwegian citizen science observations with images, and their Value of Information. The areas of the circles are relative to their respective VoI, defined as the current expected performance increase (in F_1 score) for one added observation with images for that order. If the VoI of adding data was mainly determined by the current relative over- or under-representation of a taxon, one would expect circles to gradually increase for more under-represented orders in the lower part of the graph. Numerical values provided in the Supplementary Information.

Discussion

We set out to investigate the taxonomic bias in citizen science data, in particular when accompanied by images, using a large Norwegian citizen science project

as an example case. Such images can be used to train deep neural networks for image recognition, helping citizen scientists by verifying species identifications and addressing some of the inherent taxonomic bias as they then can report within taxa they are not able to identify independently. By examining how the performance of recognition models increases as they are provided with more images in an experimental setup, we can estimate how much we expect models to improve when adding more images to those currently available for each taxon. Comparing this Value of Information (VoI) to the taxonomic bias within citizen science image data, we propose data prioritization strategies based on what additional data would improve recognition models the most. Such strategies would be more efficient than merely focusing more on taxa for which there are currently fewer images available.

Taxonomic bias

The taxonomic biases within citizen science observations considered in the current study follow a similar pattern to what has been found across biodiversity data in general[5]. However, when only considering citizen science observations with images, these trends are less pronounced; plants and fungi have relatively higher percentages of observations with images than for example birds (figure 1b). This indicates that while birds are still the most reported group also within citizen science observations with images, bird observations are generally less commonly documented with images. The reverse is true for the Insecta, which are so abundant in the citizen science image data as to be the 3rd most overrepresented class in that context. This is in stark contrast to what has been found for the totality of GBIF mediated observations globally[5] and in the Norwegian context we examined here, where the Insecta are the single most under-represented class.

Analyzing the taxonomic biases for the orders used in the machine learning part of this study sheds some light on the underlying mechanisms. While all orders within Aves are over-represented regardless of the nature of the observations considered, the Insecta are more diverse in their bias, as illustrated by Lepidoptera being the most over-represented order but Diptera the most under-represented.

We hypothesize that this disparity between taxonomic bias in all data versus that in citizen science data with images is most likely a combination of the behavior of the species and the kind of citizen scientists reporting the observations. There are distinctly different types of citizen scientists, with their own contribution patterns[34]. For casual reporters lacking specialized equipment, charismatic butterflies and flowering plants are more readily photographed opportunistically than birds. Meanwhile, a group of quite persistent ornithologists report the bulk of the bird observations in the dataset. This is typically a group reporting in a structured manner, more often based on local inventories and checklists, where reporting with images is less common than with opportunistic observations.

Image recognition and Value of Information

There are clearly differences between orders in the rates at which image recognition improves as more images are made available per species (figure 2). These differences between orders manifest in both initial performances, the rate at which performances change, and the maximum performance achieved. This indicates that, as is the case for humans, it requires more experience to learn to identify species within certain taxa than others, while the reliability with which species are correctly identified once the necessary knowledge has been acquired also differs. The differences between orders in this regard is not necessarily directly linked to the taxon's characteristics alone, however. Image quality and composition can vary between taxa depending on factors such as specimens' behavior or lack thereof, physical size, and the kind of citizen scientist generally photographing the species. A stationary flower is easier to photograph with a lot of detail than a centipede running for cover. A mite that can only be photographed with a macro lens will be photographed by a citizen scientist who has invested in such equipment. This type of citizen scientist is also more likely to invest time in taking a high quality picture than a casual citizen scientist snapping a squirrel with their mobile phone.

The VoI estimates for each of the orders provides equally diverse results. For any given number of images per species, orders differ in the expected performance increase at that point, as do the relative rates at which these performances change as data is added. As a consequence, there is a range of varying estimates for the VoI for each order, depending on both the number of images currently available per species, and the way the VoI per additional observation with images declines as more images are already available to the model.

Combining taxonomic biases and the Value of Information

We now have an estimate of how over- or under-represented the orders with which the recognition models have been trained are relative to one another, as well as a per-order estimate of the VoI per added observation with images. This means that we can address the question whether models are best improved by adding more image data equally across orders, if one should ideally prioritize under-represented orders, or if there is a prioritization to be made based on order-specific differences. As shown in figure 4, there are distinct differences in the VoI per order, and these do not merely correlate with their respective over- or under-representation. The plant orders of Asparagales and Lamiales clearly have a higher VoI despite their slight over-representation when compared to the other orders in this experiment. The fungi order Polyporales also gains more than twice the VoI per additional observation with images in comparison to the fourth-most valuable order, the Lecanorales. We conclude that, from a VoI perspective, these are the orders for which a recognition model would benefit the most per observation with images added, despite the fact that other orders are numerically more under-represented.

Conclusions

Based on the Value of Information (VoI) for image recognition models, a citizen scientist or citizen science project manager aiming to maximize their impact in this regard might want to focus on orders with the highest expected VoI per observation with images added, rather than simply on the order with the lowest number of images per species. Observations with images of other orders, while in some cases less well-represented in the available image data, appear to provide less VoI per additional observation. As citizen scientists are in large part motivated by a desire to advance scientific knowledge[35], communicating such considerations can be an important part of community engagement.

In generalizing these findings, the following has to be noted:

- The taxa identified here as having the highest expected VoI per observation with images added are examples from the limited subset of orders used within this experiment. As illustrated by the observed variation in per-species representation and VoI between orders that belong to the same class, it is evident that generalization of a class like Insecta fails to give insight into intra-class variation. It is likely that a similar principle applies to orders, where for example a taxonomic group like Norwegian warblers likely has a different VoI curve than the more readily distinguished titmice. Such differences will remain hidden from view when analyzing passerine birds as a single taxonomic group.

- Our findings are derived from Norwegian species reported on a single Norwegian citizen science portal. The diversity of species within the same orders can differ in other regions, affecting the VoI curves. Different portals will also differ in the way they accommodate reporting observations with images, and in general attract different types of users[23]. All of these factors are likely to have an effect on the proportion of observations accompanied by photographic evidence and the quality thereof. Such factors also affect the nature of newly added data, including its expected VoI.

- Models were trained on species for which at least 220 observations with images were available. This is not a random subset of all the species within an order, and likely to be biased towards charismatic species and those that are more readily identifiable from an image. This can lead to an overestimation in terms of learning rate and thus the VoI curve, especially within orders in which relatively few species have the data availability we selected for here. Then again, future observations to be added to the data will be prone to the same biases, in which case the VoI of such an addition will be lower than it would be for a truly random species.

- Current and future (deep) learning methods alternative to Convolutional Neural Networks (CNN) may be able to utilize more information in an image

and generalize more rapidly, using less data. This could have implications for the importance of VoI relative to the overall bias. We expect that the demonstrated differences in VoI between species are not unique to CNN however, and in part inherent to the visual information available in each picture. Either way, awareness of the potential differences in VoI between taxa is warranted, and an interesting consideration to evaluate in future studies.

Regardless of the specific taxa and derived values, our findings demonstrate that a more informed decision is possible when choosing to focus on certain taxa for data collection aimed at improved recognition models. Prioritization of taxa for which to mobilize additional data can be informed by considering its expected VoI, rather than simply prioritizing those that are currently the most under-represented numerically. Note that this is no plea for deprioritizing data collection for such taxa in the context of citizen science as a whole. There are many areas of management and research that can benefit from additional data on taxa we predict will benefit less from additional images for recognition models, and ample reasons to mobilize data for other applications than image recognition.

Training machine learning models requires substantial amounts of data, certainly when context, morphology and phenology vary, such as when classifying *in situ* images. Data collection in machine learning generally is a matter of harvesting whatever one can to provide the model with more data. Within (citizen) science, the collection of images mainly serves as secondary data, providing documentation for the occurrence it accompanies. With the more widespread use of image recognition models as both a user tool and a mechanism for quality control, it is time to view images as data in and of themselves. Such a shift calls not only for conscious choices when it comes to the VoI in images, but increased implementation of data practices such as persistent storage, metadata standardization and the other FAIR data principles[36] to enable more apt usage of image data for current and novel applications.

Methods

In the current study we utilize an extensive network and data from citizen science in order to test for among taxa variation in biases and Value of Information (VoI) in image recognition training data. We use data from the Norwegian Species Observation Service as an example dataset due to the generic nature of this citizen science platform, where all multicellular taxa from any Norwegian region can be reported both with and without images. The platform is open to anyone willing to report under their full real name, and does not record users' expertise or profession. The platform had 6,205 active contributors in 2021 out of its 17,655 registered users, and currently publishes almost 27 million observations through GBIF, of which 1.08 million with one or more images. Observations have been bulk-verified by experts

appointed by biological societies receiving funding for this task, with particular focus on red listed species, invasive alien species, and observations out of range or season. Observations containing pictures receive additional scrutiny, as other users can alert reporters and validators to possible mistaken identifications. An advantage of this particular platform is that no image recognition model has been integrated. This ensures that the models trained in this experiment are not trained on the output resulting from the use of any model, but with identifications and taxonomic biases springing from the knowledge and interest of human observers. Moreover, the platform's compliance with the authoritative Norwegian taxonomy allows for analyses on taxonomic coverage.

In an exploration procedure we determined the taxonomic level of orders to be suitable examples of taxa with a sufficiently wide taxonomic diversity, and enough data in the dataset to be evaluated for models in this experiment. Data collection was done by acquiring taxon statistics and observation data from the Global Biodiversity Information Facility (GBIF), the largest aggregator of biodiversity observations in the world[37] for the selected orders, as well as the classes used by Troudet *et al.*[5]. The authoritative taxonomy for Norway was downloaded from the Norwegian Biodiversity Information Centre[38]. In the experimental procedure, models were trained for 12 distinct orders (listed in figure 4), artificially restricting these models to different amounts of data. In the data analysis stage, model performances relative to the amount of training data were fitted for each order, allowing the estimation of a VoI. Using the number of observations per species on GBIF, and the number of species known to be present in Norway from the Norwegian Species Nomenclature Database, we calculated relative taxonomic biases.

Exploration

Initial pilot runs were done on 8 taxa (see Supplementary Information), using different subset sizes of observations for each species, and training using both an Inception-ResNet-v2[39] as well as an EfficientNetB3[40] architecture for each of these subsets. These initial results indicated that the Inception-ResNet-v2 performance (F_1) varied less between replicate runs and was generally higher, so subsequent experiments were done using this architecture. The number of observations which still improved the accuracy of the model was found to be between 150 and 200 in the most extreme cases, so the availability of at least 220 observations with images per species was chosen as an inclusion criteria for the further experiment. This enabled us to set aside at least 20 observations per species as a test dataset for independent model analysis.

From a Darwin Core Archive file of Norwegian citizen science observations from the Species Observation Service with at least one image[33], a tally of the number of such observations per species was generated. We then calculated how many species, with a minimum of 220 such observations, would, at a minimum, be

available per taxon if a grouping was made based on each taxon rank level with the constraint of resulting in at least 12 distinct taxa. For each taxonomic level, we calculated how many species having at least 220 such observations were available per taxon when dividing species based on that taxon level. When deciding on the appropriate taxon level to use, we limited the options to taxon levels resulting in at least 12 different taxa.

A division by order was found to provide the highest minimum number of species (17) per order within these constraints, covering 12 of the 96 eligible orders. The next best alternative was the family level, which would contain 15 species per family, covering 12 of the 267 eligible families.

Data collection

We retrieved the number of species represented in the Norwegian data through the GBIF API, for all observations, all citizen science observations, and all citizen science observations with images for the 12 selected orders and the classes used by Troudet *et al.*[5]. We also downloaded the Norwegian Species Nomenclature Database[38] for all kingdoms containing taxa included in these datasets. Observations with images were collected from the Darwin Core Archive file used in the exploration phase, filtering on the selected orders. For these orders, all images were downloaded and stored locally. The average number of images per observation in this dataset was 1.44, with a maximum of 17 and a median of 1.

Experimental procedure

For each selected order, a list of all species with at least 220 observations with images was generated from the Darwin Core Archive file[33]. Then, runs were generated according to the following protocol (figure 5):

1. From a list sorted alphabetically by the full taxonomy of the species, a subset of 17 consecutive species starting from a random index was selected. If the end of the list was reached with fewer than 17 species selected, selection continued from the start of the list. The taxonomic sorting ensures that closely related species (belonging to the same family or genus), bearing more similarity, are more likely to be part of the same experimental set. This ensures that the classification task is not simplified for taxa with many eligible species.

2. Each of the 220+ observations for each species were tagged as being either test, training or validation data. A random subset of all but 200 were assigned to the test set. The remaining 200 observations were, in a 9:1 ratio, randomly designated as training or validation data, respectively. In all cases, images from the same observation were assigned to the same subset, to keep

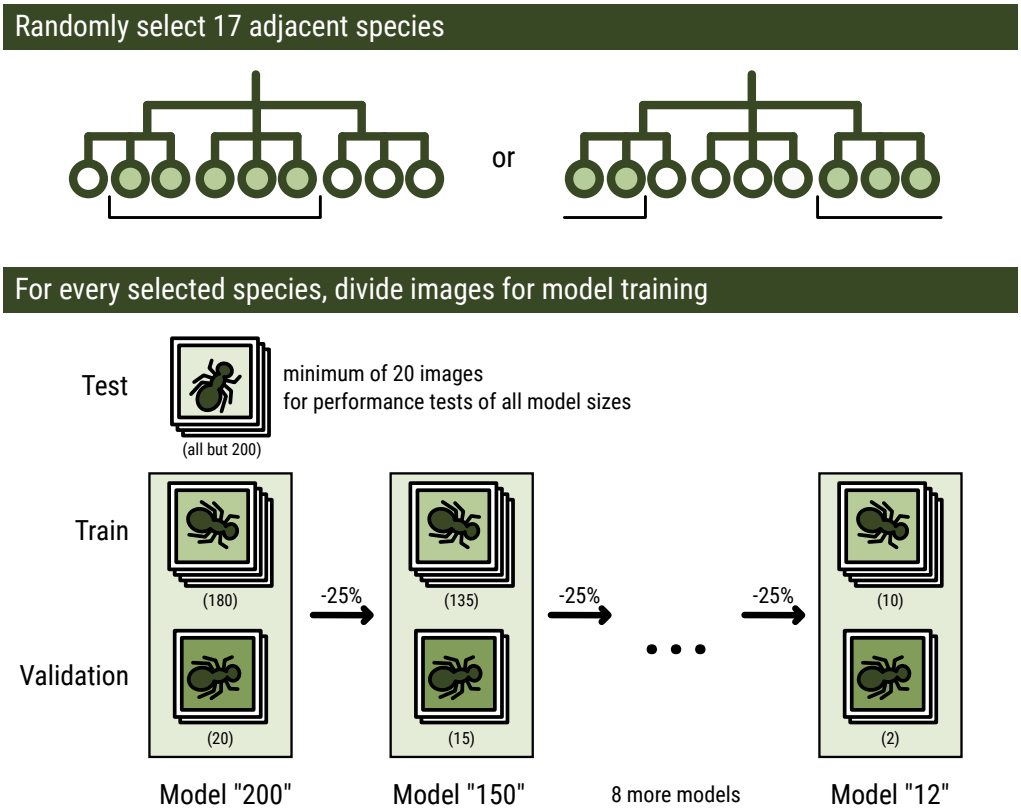


Figure 5: Data selection and subdivision. Each run is generated by selecting 17 taxonomically adjacent species per order, and randomly assigning all available images of each selected species to that run's test-, train- or validation set. Training data are used as input during training, using the validation data to evaluate performance after each training round in order to adjust training parameters during training. The test set is used to measure model performance independently after the model is finalized[28]. For each subsequent model in that run, training and validation data are reduced by 25% (or slightly less than 25% if not divisible by 4). The test set is not reduced, and used for all models within a run.

the information in each subset independent from the others. The resulting lists of images are stored as the test set and 200-observation task.

3. The 200 observations in the training and validation sets were then repeatedly reduced by discarding a random subset of 25% of both, maintaining a validation data proportion of $\leq$10%. The resulting set was saved as the next task, and this step was repeated as long as the resulting task contained a minimum of 10 observations per species. The test set remained unaltered throughout.

Following this protocol results in a single run of related training tasks with 200, 150, 113, 85, 64, 48, 36, 27, 21, 16 and 12 observations for training and

validation per species. The seeds for the randomization for both the selection of the species and for the subsetting of training- and validation datasets were stored for reproducibility. The generation of runs was repeated 5 times per order to generate runs containing tasks with different species subsets and different observation subsetting.

Then, a Convolutional Neural Network based on Inception-ResNet-v2[39] (see the Supplementary Information for model configuration) was trained using each predesignated training/validation split. When the learning rate had reached its minimum and accuracy no longer improved on the validation data, training was stopped and the best performing model was saved. Following this protocol, each of the 12 orders were trained in 5 separate runs containing 11 training tasks each, thus producing a total of 660 recognition models. After training, each model was tested on all available test images for the relevant run.

Data analysis

The relative representation of species within different taxa were generated using the number of species present in the GBIF data for Norway within each taxon and the number of accepted species within that taxon present in the Norwegian Species Nomenclature Database[38], in line with Troudet *et al.*[5]: $R_x = n_x - (n\frac{s_x}{s})$ where R_x is the relative representation for taxon x, n_x is the number of observations for taxon x, n is the total number of observations for all taxa, s_x is the number of species within taxon x, and s is the total number of species within all taxa.

As a measure of model performance, we use the F_1 score, the harmonic mean of the model's precision and recall, given by

$$F_1 = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

where tp, fp and fn stand for true positives, false positives and false negatives, respectively. The F_1 score is a commonly used metric for model evaluation, as it is less susceptible to data imbalance than model accuracy[28].

The Value of Information (VoI) can be generically defined as "*the increase in expected value that arises from making the best choice with the benefit of a piece of information compared to the best choice without the benefit of that same information*"[32]. In the current context, we define the VoI as the expected increase in model performance (F_1 score) when adding one observation with at least one image. To estimate this, for every order included in the experiment, the increase in average F_1 score over increasing training task sizes were fitted using the Von Bertalanffy Growth Function, given by

$$L = L_\infty(1 - e^{-k(t-t_0)}).$$

where L is the average F_1 score, L_∞ is the asymptotic maximum F_1 score, k is the growth rate, t is the number of observations per species, and t_0 is a hypothetical

number of observations at which the F_1 score is 0. The Von Bertalanffy curve was chosen as it contains a limited number of parameters which are intuitive to interpret, and fits the growth of model performance well.

The estimated increase in performance at any given point is then given by the slope of this function, i.e. the result of the differentiation of the Von Bertalanffy Growth Curve, given[41] by

$$\frac{dL}{dt} = bke^{-kt}$$

where

$$b = L_\infty e^{kt_0}.$$

Using this derivative function, we can estimate the expected performance increase stemming from one additional observation with images for each of the species within the order. Filling in the average number of citizen science observations with images per Norwegian species in that order for t, and dividing the result by the total number of Norwegian species within the order, provides the VoI of one additional observation with images for that order, expressed as an average expected F_1 increase.

References

1. Xu, H. *et al.* Ensuring effective implementation of the post-2020 global biodiversity targets. *Nature Ecology & Evolution* **5,** 411–418. https://doi.org/10.1038/s41559-020-01375-y (Jan. 2021).
2. Pereira, H. M. *et al.* Essential Biodiversity Variables. *Science* **339,** 277–278. https://doi.org/10.1126/science.1229931 (Jan. 2013).
3. Rocha-Ortega, M., Rodriguez, P. & Córdoba-Aguilar, A. Geographical, temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. *Ecological Entomology* **46,** 718–728. https://doi.org/10.1111/een.13027 (Feb. 2021).
4. Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* **19,** 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002 (Jan. 2014).
5. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7.** https://doi.org/10.1038/s41598-017-09084-6 (Aug. 2017).
6. GBIF.org. *Global data trends* https://www.gbif.org/analytics/global.

7. Bertacchi, A., Giannini, V., Franco, C. D. & Silvestri, N. Using unmanned aerial vehicles for vegetation mapping and identification of botanical species in wetlands. *Landscape and Ecological Engineering* **15,** 231–240. https://doi.org/10.1007/s11355-018-00368-1 (Feb. 2019).
8. Tollefson, J. Computers on the reef. *Nature* **537,** 123–124. https://doi.org/10.1038/537123a (Aug. 2016).
9. August, T. *et al.* Emerging technologies for biological recording. *Biological Journal of the Linnean Society* **115,** 731–749. https://doi.org/10.1111/bij.12534 (Apr. 2015).
10. Christin, S., Hervet, É. & Lecomte, N. Applications for deep learning in ecology. *Methods in Ecology and Evolution* **10** (ed Ye, H.) 1632–1644. https://doi.org/10.1111/2041-210x.13256 (July 2019).
11. Silvertown, J. A new dawn for citizen science. *Trends in Ecology & Evolution* **24,** 467–471. https://doi.org/10.1016/j.tree.2009.03.017 (Sept. 2009).
12. Chandler, M. *et al.* Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* **213,** 280–294. https://doi.org/10.1016/j.biocon.2016.09.004 (Sept. 2017).
13. Theobald, E. *et al.* Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation* **181,** 236–244. https://doi.org/10.1016/j.biocon.2014.10.021 (Jan. 2015).
14. Pocock, M. J. *et al.* in *Advances in Ecological Research* 169–223 (Elsevier, 2018). https://doi.org/10.1016/bs.aecr.2018.06.003.
15. Chandler, M. *et al.* in *The GEO Handbook on Biodiversity Observation Networks* (eds Walters, M. & Scholes, R. J.) 211–237 (Springer International Publishing, Cham, 2017). ISBN: 978-3-319-27288-7. https://doi.org/10.1007/978-3-319-27288-7_9.
16. Trouille, L., Lintott, C. J. & Fortson, L. F. Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proceedings of the National Academy of Sciences* **116,** 1902–1909. https://doi.org/10.1073/pnas.1807190116 (Feb. 2019).
17. Bonney, R., Phillips, T. B., Ballard, H. L. & Enck, J. W. Can citizen science enhance public understanding of science? *Public Understanding of Science* **25,** 2–16. https://doi.org/10.1177/0963662515607406 (Oct. 2015).
18. Schuttler, S. G., Sorensen, A. E., Jordan, R. C., Cooper, C. & Shwartz, A. Bridging the nature gap: can citizen science reverse the extinction of experience? *Frontiers in Ecology and the Environment* **16,** 405–411. https://doi.org/10.1002/fee.1826 (July 2018).

19. Crall, A. W. *et al.* Assessing citizen science data quality: an invasive species case study. *Conservation Letters* **4,** 433–442. https://doi.org/10.1111/j.1755-263x.2011.00196.x (Aug. 2011).

20. Burgess, H. *et al.* The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation* **208,** 113–120. https://doi.org/10.1016/j.biocon.2016.05.014 (Apr. 2017).

21. Callaghan, C. T. *et al.* Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience* **71,** 55–63. ISSN: 0006-3568. https://doi.org/10.1093/biosci/biaa131 (Nov. 2020).

22. Bayraktarov, E. *et al.* Do Big Unstructured Biodiversity Data Mean More Knowledge? *Frontiers in Ecology and Evolution* **6.** https://doi.org/10.3389/fevo.2018.00239 (Jan. 2019).

23. Boakes, E. H. *et al.* Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific Reports* **6.** https://doi.org/10.1038/srep33051 (Sept. 2016).

24. Weinstein, B. G. A computer vision for animal ecology. *Journal of Animal Ecology* **87** (ed Prugh, L.) 533–545. https://doi.org/10.1111/1365-2656.12780 (Nov. 2017).

25. Wäldchen, J., Rzanny, M., Seeland, M. & Mäder, P. Automated plant species identification—Trends and future directions. *PLOS Computational Biology* **14** (ed Bucksch, A.) e1005993. https://doi.org/10.1371/journal.pcbi.1005993 (Apr. 2018).

26. Ceccaroni, L. *et al.* Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. *Citizen Science: Theory and Practice* **4.** https://doi.org/10.5334/cstp.241 (2019).

27. Wäldchen, J. & Mäder, P. Machine learning for image based species identification. *Methods in Ecology and Evolution* **9** (ed Cooper, N.) 2216–2225. https://doi.org/10.1111/2041-210x.13075 (Sept. 2018).

28. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org. http://www.deeplearningbook.org (MIT Press, 2016).

29. Seltzer, C., Ueda, K.-i. & Shepard, A. *A New Vision Model!* https://www.inaturalist.org/blog/31806-a-new-vision-model.

30. Terry, J. C. D., Roy, H. E. & August, T. A. Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution* **11** (ed Altwegg, R.) 303–315. https://doi.org/10.1111/2041-210x.13335 (Jan. 2020).

31. Horn, G. V. *et al. The iNaturalist Species Classification and Detection Dataset* 2018. arXiv: 1707.06642 [cs.CV]. https://doi.org/10.48550/arXiv.1707.06642.

32. Keisler, J. M., Collier, Z. A., Chu, E., Sinatra, N. & Linkov, I. Value of information analysis: the state of application. *Environment Systems and Decisions* **34,** 3–23. https://doi.org/10.1007/s10669-013-9439-4 (Apr. 2013).

33. GBIF.org. *GBIF Occurrence Download* 2021. https://doi.org/10.15468/dl.tc4w55.

34. Aristeidou, M., Scanlon, E. & Sharples, M. Profiles of engagement in online communities of citizen science participation. *Computers in Human Behavior* **74,** 246–256. https://doi.org/10.1016/j.chb.2017.04.044 (Sept. 2017).

35. Richter, A. *et al.* Motivation and support services in citizen science insect monitoring: A cross-country study. *Biological Conservation* **263,** 109325. https://doi.org/10.1016/j.biocon.2021.109325 (Nov. 2021).

36. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3.** https://doi.org/10.1038/sdata.2016.18 (Mar. 2016).

37. GBIF.org. *GBIF homepage* https://www.gbif.org.

38. Norwegian Biodiversity Information Centre. *Species Nomenclature Database* http://eksport.artsdatabanken.no/Artsnavnebase (2021).

39. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning* 2016. arXiv: 1602.07261 [cs.CV]. https://doi.org/10.48550/arXiv.1602.07261.

40. Tan, M. & Le, Q. V. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* 2020. arXiv: 1905.11946 [cs.LG]. https://doi.org/10.48550/arXiv.1905.11946.

41. Campbell, N. A. & Phillips, B. F. The Von Bertalanffy Growth Curve and Its Application to Capture – Recapture Data in Fisheries Biology. *ICES Journal of Marine Science* **34,** 295–299. https://doi.org/10.1093/icesjms/34.2.295 (Mar. 1972).

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the GBIF repository, https://doi.org/10.15468/dl.tc4w55.

All code used in this study for the experiment and the generation of this manuscript and its graphs is available on https://github.com/WouterKoch/citizen_science_VoI.

Maximizing citizen scientists' contribution to automated species recognition
SUPPLEMENTARY MATERIALS

Pilot run taxa

Initial pilot runs were done on Bombus, Cetoniidae, Coccinellidae, Coleoptera, Lepidoptera, Odonata, Rodentia, and Zygaenidae. These taxa were chosen in order to test the machine learning pipeline on taxa with different levels of difficulty of identification. Rodentia were added to include a taxon outside of the Insecta.

Image recognition model configuration

Models were trained in Python 3.9[1], using TensorFlow[2] and Keras[3] to train a new recognition model based on the Inception-ResNet-v2 architecture[4] for every dataset. A dense classification layer using softmax activation replaced the top layer of the Inception-ResNet-v2 model as a new top layer, with 17 nodes to classify each of the 17 species. For the loss function we used standard categorical cross entropy loss.

Color channels of input images were normalized between -1 and 1, and were scaled to 256×256 pixels, cropping the image to become square if needed. Training data were augmented by shearing up to a factor of 0.2, zooming up to a factor of 0.2, rotating up to 90 degrees, and randomly flipping horizontally or not. Validation and test images were only normalized and squared, not augmented.

In the first training stage, the weights of the original Inception-ResNet-v2 layers were frozen, training only the newly added top layer. This was done for 2 epochs with a learning rate of $1 \cdot 10^{-3}$. This has an equivalent effect as learning rate warm-up.

In the second training stage, all layers were trained. This was done for a maximum of 200 epochs, with an initial learning rate of $1 \cdot 10^{-4}$. The learning rate was multiplied by 0.1 when the validation loss did not improve for 3 consecutive epochs. The minimum of the learning rate was set to $1 \cdot 10^{-8}$.

After each epoch, model performance was evaluated using the validation set, saving the weights of the current model to disk as the latest checkpoint if the accuracy for the validation set had improved since the last saved checkpoint. Finally, when the model did not reduce its loss for 8 consecutive epochs, training was stopped. The most recently stored checkpoint was then used as the final recognition model for that dataset, and its performance measured using the test data.

Taxonomic order result metrics

Order	Bias in cs data with img	VoI (F_1 increase $\cdot 10^6$)
Asparagales	5259	13.05
Lamiales	3879	9.36
Polyporales	-11060	4.11
Lecanorales	-106853	1.72
Agaricales	-22932	1.52
Diptera	-110248	1.42
Coleoptera	-51782	1.09
Passeriformes	28630	0.73
Odonata	13075	0.32
Lepidoptera	145110	0.17
Charadriiformes	57421	0.13
Anseriformes	49501	0.02

Table S1: Orders used in the machine learning experiment, their over- or under-representation among citizen science observations with images (relative to all orders having an equal average amount of such observations per species), and the Value of Information as measured by the expected F_1 increase for adding one observation with images to the number of observations with images currently available. Sorted by VoI (descending). These are the numerical values for figure 4.

Von Bertalanffy Growth Curves

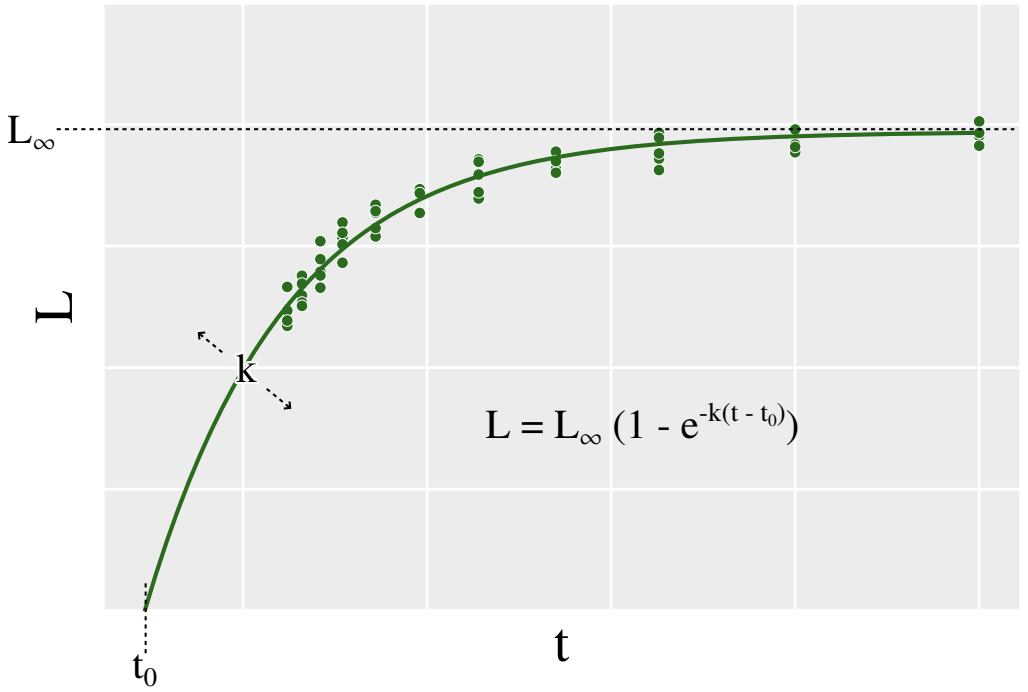


Figure S1: Visualization of the Von Bertalanffy Growth Curve parameters. Curves were fitted using the Levenberg-Marquardt (Least Squares) algorithm. Residuals were plotted for each taxon and not found to be heterogeneous in their distribution.

References

1. Python Software Foundation. *Python Language Reference, version 3.9* `http://www.python.org`.
2. Martín Abadi *et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* Software available from tensorflow.org. 2015. `https://www.tensorflow.org/`.
3. Chollet, F. *et al. Keras* 2015. `https://keras.io`.
4. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning* 2016. arXiv: `1602.07261 [cs.CV]`. `https://doi.org/10.48550/arXiv.1602.07261`.

Paper III

Recognizability bias in citizen science photographs

Wouter Koch[1,2] Laurens Hogeweg[3,4] Erlend B. Nilsen[5]
Robert B. O'Hara[6] Anders G. Finstad[1]

1: Department of Natural History, Norwegian University of Science and Technology, Erling Skakkes gate 47b, Trondheim, Norway; 2: Norwegian Biodiversity Information Centre, Havnegata 9, 7010 Trondheim, Norway; 3: Intel Benelux, High Tech Campus 83, 5656 AE Eindhoven, The Netherlands; 4: Naturalis Biodiversity Center, PO Box 9517, 2300 RA, Leiden, The Netherlands; 5: Norwegian Institute for Nature Research, Postboks 5685 Torgarden, 7485 Trondheim, Norway; 6: Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, 7034 Trondheim, Norway

Preprint on bioRχiv: doi:10.1101/2022.06.25.497604

Abstract

Citizen science initiatives and automated collection methods increasingly depend on image recognition in order to provide the amounts of observational data research and management needs. Training recognition models, meanwhile, also requires large amounts of data from these sources, creating a feedback loop between the methods and the tools. Species that are harder to recognize, both for humans and machine learning algorithms, are likely to be underreported, and thus be less prevalent in the training data. As a result, the feedback loop may hamper training mostly for species that already pose the greatest challenge. In this study, we trained recognition models for various taxa, and found evidence for a *"recognizability bias"*, where species that models struggle with are also generally underreported. This has implications for the kind of performance one can expect from future models that are trained with more data, including such challenging species. We consider identification methods that rely on more than photographs alone to be important in improving future identification tools.

Introduction

There is an ever growing need for large amounts of biodiversity observation data. With an increasing awareness of the multiple crises biodiversity faces[1–3],

substantial amounts of such data are essential if humanity is to monitor trends and address these issues[4–6]. Occurrence data are typically subject to spatial, temporal and taxonomic bias[7,8], and traditional manual methods of data collection are insufficient to gather the data volume needed, or address these biases. Alternative data collection methods, ranging from citizen science (non-professional volunteers reporting observations[9]) to camera-traps automating insect monitoring[10,11] are being deployed to gather large amounts of data. With the increased output from such initiatives, manual management and quality control become infeasible. Automated image recognition tools for species identification are increasingly used to facilitate this[12–15]. Training image recognition models, however, also requires large amounts of pictures[16]. This creates a mutual reliance between large scale image data collection and image recognition models[17].

Visual identification of species is a complex task, and taxa vary in their recognizability; while some species are unmistakable, many others are very challenging or even outright impossible to identify, regardless of picture quality[18]. As models are trained using training data reported and identified by humans, species with low recognizability among humans will be underreported and be underrepresented in the training data. This affects recognition models, as these are then being trained with data biased towards higher recognizability, consisting mostly of pictures of species that are easier to recognize. If this is the case, training models will be hampered not only by the lower recognizability of particularly challenging species, but also by their higher absence from the training data.

To evaluate the existence of this possible bias and its consequences, we evaluated how data availability, picture quality, biological traits and data collection differs across species within 3 orders of birds, and how these differences relate to recognition model performance. All data came from a large Norwegian citizen science project, where recognition tools are not a part of the reporting or validation process. Birds are the most well-represented orders per species, allowing for the most detailed analysis. We also trained models for 9 other orders of plants, animals and fungi, to test for a general correlation between data availability and model performance, and to evaluate what this means for future recognition models.

We find evidence for a *"recognizability bias"*, where species that are more readily identified by humans and recognition models alike are more prevalent in the available image data. This pattern is present across multiple taxa, and does not appear to relate to a difference in picture quality, biological traits, or data collection metrics other than recognizability.

Methods

We trained image recognition models using convolutional neural networks on pictures retrieved from the Norwegian citizen science platform Species Observation Service[19] for 12 orders: Agaricales, Anseriformes, Asparagales, Asterales, Charadri-

iformes, Coleoptera, Diptera, Lecanorales, Lepidoptera, Odonata, Passeriformes, and Polyporales[20]. A separate model was trained for each order, using 200 documented observations per species for training and validation, and a minimum of 20 for the test set. See Koch *et al.*[21] for details. From these models and various external datasets, several relevant metrics were collected (table 1).

Metric	**Definition**
Data availability	The total number of citizen science observations from the Norwegian citizen science platform Species Observation Service[19] for a species, containing one or more pictures. This is a more meaningful measure than simply the total number of pictures, as multiple pictures within an observation are not independent from one another and therefore do not add as much information as unique observations.
F_1-score	The performance obtained for a species in a recognition model, defined as the harmonic mean of the precision and recall[16]
Species in Norway	The number of species within an order that are present in Norway, according to the Norwegian Species Nomenclature Database[22].

Table 1: Metrics collected for species within all orders

More detailed analyses were done on the included bird orders; waterfowl (Anseriformes), shorebirds (Charadriiformes), and passerines (Passeriformes), as bird orders have the highest proportion of species in Norway represented in the dataset, and ample standardized available data on a range of biological traits allowing for a deeper analysis. For these analyses, a number of additional metrics were collected for the included bird species (table 2).

Metric	**Definition**
Picture quality	Using Label Studio v1.4[23], $\geq$50 pictures per species were annotated by drawing rectangles approximately equal in surface area to the visible part of each individual bird. From this, we took the percentage of the picture occupied by the largest depiction of an individual of the target species, minus the percentage of the picture occupied by all individuals of other bird species. Per species, the median log value was used as a proxy for picture quality.
Urbanness	The proportion of 100 documented observations from the Species Observation Service with a location within a cell tagged as "urban" in the ESA CCI landcover dataset[24].

Hand-wing index	Wing length minus wing width, a measure positively correlated with flight efficiency and dispersal ability of a species. Retrieved from the Global-HWI dataset[25].
Body mass	The average log-transformed body mass of a species, retrieved from the Global-HWI dataset[25].
Habitat openness	A three-step scale of the openness of the habitat of a species, retrieved from the Global-HWI dataset[25].
Documentation rate	The proportion, per species, of observations in the Species Observation Service that have one or more pictures.
Picture density	The average number of pictures per observation from the Species Observation Service, from those with at least one picture.
Observation rate	The number of observations in the Species Observation Service dataset per observation in the TOV-e bird monitoring scheme[26]

Table 2: Metrics collected for species within the bird orders

LASSO multiple regression models were trained using Scikit-learn[27] to evaluate the effect of the biological traits, picture quality measurement, and data collection process from table 2 on the F_1-scores for birds. All LASSO models have the order as a factor. The full model for biological traits is given by

$$F_1 = \beta_0 + \beta_1 HWI + \beta_2 BM + \beta_3 H + \beta_4 U + \beta_5 DA + \epsilon + (1|Order)$$

where HWI is the hand-wing index, BM is the body mass, H is the habitat openness, U is the urbanness, and DA is the log data availability. The full model for picture quality is given by

$$F_1 = \beta_0 + \beta_1 Q + \beta_2 DA + \epsilon + (1|Order)$$

where Q is the picture quality, and DA is the log data availability. The full model for data collection parameters is given by

$$F_1 = \beta_0 + \beta_1 OR + \beta_2 DR + \beta_3 PD + \beta_4 DA + \epsilon + (1|Order)$$

where OR is the observation rate, DR is the documentation rate, PD is the picture density, and DA is the log data availability.

Results

There is a strong positive linear correlation between log data availability and the F_1-score for bird species (figure 1). Note that data availability does not affect training,

as all models were trained and evaluated using 220 documented observations per species, regardless of the total availability. A positive linear correlation was also evident in 7 of the 9 other orders (figure 2), in particular Asterales and Odonata. The beetles (Coleoptera) and lichens (Lecanorales) exhibited no apparent correlation, with an R^2 of 0.06 and 0.12, and P-values of 0.27 and 0.18, respectively.

In each bird order, there is a linear relationship between species' picture density and documentation rate ($R^2 \geq 0.52$, $p \leq 1.51\times10^{-7}$, see table S2). We also find a negative linear correlation between picture density and F_1-scores ($R^2 \geq 0.23$, $p \leq 2.1\times10^{-4}$, see table S2), and some negative linear correlation between documentation rate and F_1-scores ($R^2 \geq 0.11$, $p \leq 4.64\times10^{-3}$, see table S2). For passerines, there is a negative linear relationship between habitat openness and picture quality ($R^2 = 0.26$, $p = 3.53\times10^{-8}$, see table S2). Waterfowl and shorebirds could not be evaluated as they only occur in open habitats.

LASSO models trained on biological traits, collection process parameters, and picture quality, all having and log data availability as an additional parameter and order as a factor, had R^2 values of 0.60, 0.57 and 0.63, respectively. With that, none of the full model performances were substantial improvements from a LASSO model with log data availability as its only parameter ($R^2 = 0.57$).

Discussion

We find a conspicuous pattern where recognition models attain higher performances for species that are reported with pictures more frequently. It is probable that the recognizability of the species influences both their likelihood of being reported with pictures, as well as recognition model performances. The citizen science project used as a data source here does not include any recognition tools in its reporting or validation process, allowing a distinction between human and algorithm recognition biases. Unmistakable species can be recognized and reported by more citizen scientists, resulting in greater data availability for such species. A recognition model, dealing with the same information as human observers, is also proportionally more likely to reliably recognize these species.

This is supported by a qualitative comparison between species with the highest and lowest recognition model performances, where easy to recognize, characteristic species are reported more often than hard to recognize species (e.g. nondescript species or species similar to other related species) (see figure 1 and table S1). Further support comes from the fact that most of the correlation is explained by the data availability for a species, rather than the documentation rate or the picture density. Thus, there is more data available mainly when a species is recognized and reported more, rather than it being disproportionately more likely to be reported with pictures, or with many pictures when reported with pictures.

An alternative explanation to recognizability for increased model performance

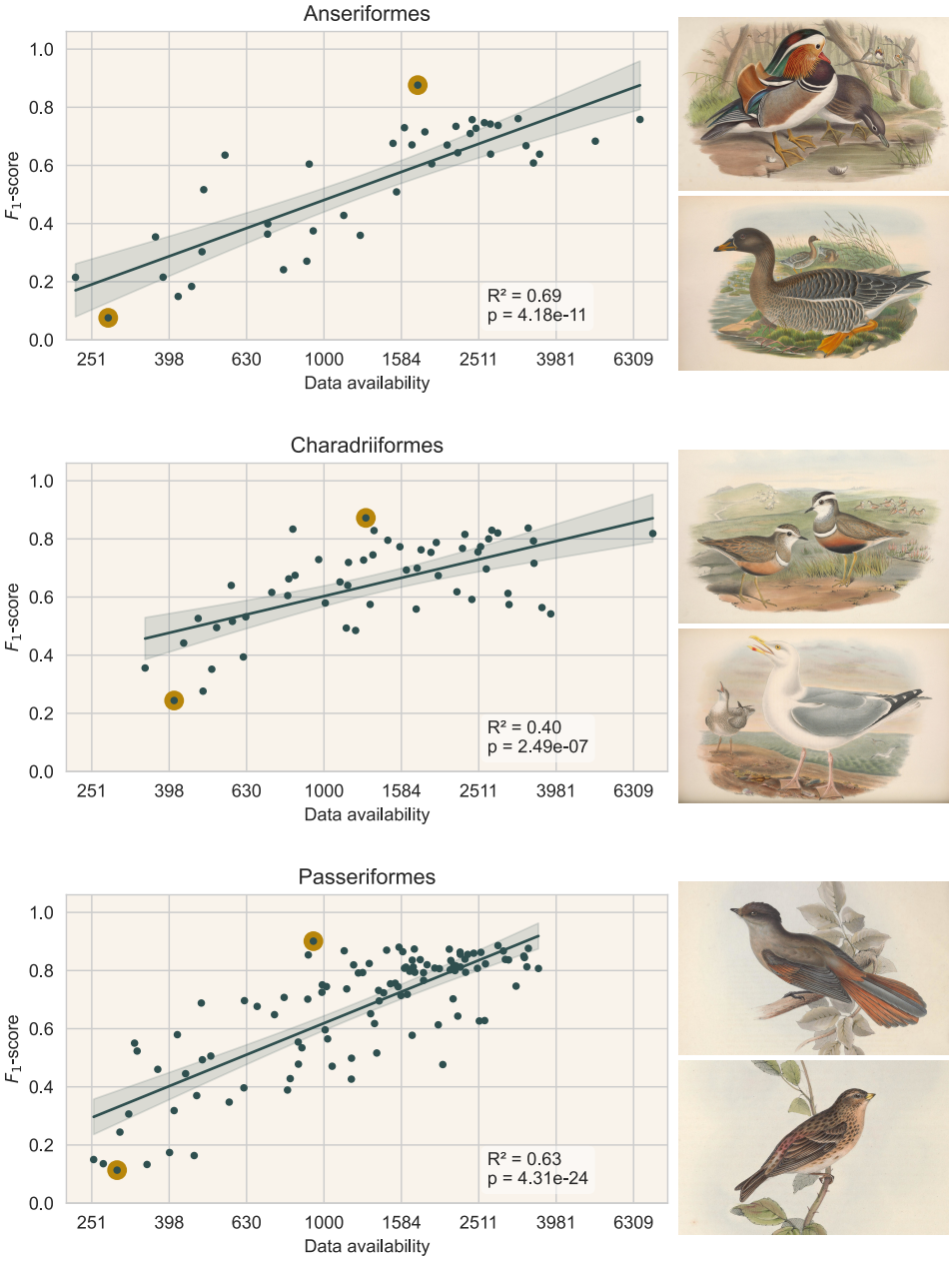


Figure 1: Effect of the total data availability per species on their F_1-scores, in models trained with 200 documented observations, for three bird orders. The top- and bottom-performing species per order (highlighted dots) are depicted, see table S1. Regressions are Ordinary Least Squares with 95% confidence intervals.

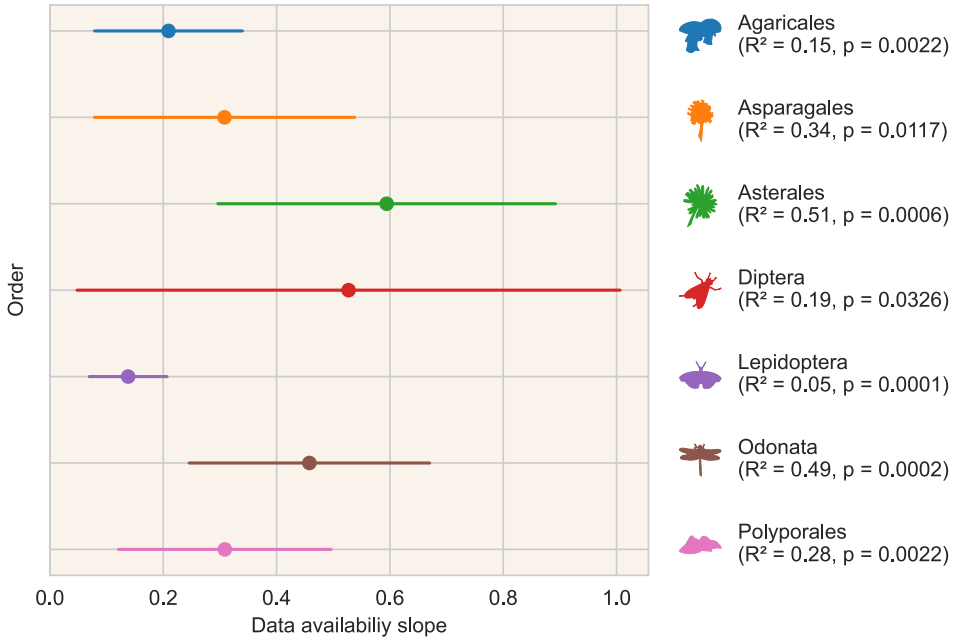


Figure 2: The slopes of the correlations between total data availability per species and their F_1-scores, in models trained with 200 documented observations, for non-bird orders with a correlation $p<0.05$. Regressions are Ordinary Least Squares, lines indicate the 95% confidence intervals.

might be a difference in the kind of pictures, but we find no evidence for this. Species traits, habitat use, and image quality could affect recognition model performance if pictures of more photographed birds are taken more up close, with higher zoom, or were cropped more. We found no evidence, however, for a link between model performance and either picture quality or biological traits in birds. For the passerines, where habitat openness varies among species, we do find that picture quality decreases for species associated with more open habitats. It makes intuitive sense that birds in open habitats are photographed from a greater distance than their forest dwelling counterparts, which will be hidden from view unless in close proximity. While this intercorrelation supports the validity of the picture quality metric, neither habitat nor picture quality affect recognition model performance. We conclude that differences in model performance are caused by the recognizability of the species, rather than by how, or how large species are generally depicted.

Since multiple pictures connected to a single observation are not truly independent, training data are generated based on the number of documented observations, rather than the total number of pictures. One might expect that species with a higher picture density will perform better, as observations with more pictures can

provide some additional information in the training process. We find a reverse effect however, where performance for such species is substantially lower. A likely explanation is that species with high picture densities are rarities in Norway (e.g. the top 3 species being Caspian gull, Blyth's reed warbler, and Pine bunting). Species with the lowest picture density, meanwhile, are typical common, well-known species such as corvids and titmice. Rarities are reported not because they are easy to find or identify by casual observers, but due to their popularity among avid birdwatchers, who are likely to document their observations. A strong correlation between picture density and documentation rate supports this; rarities are more often reported with pictures, and in such cases relatively often with several pictures.

While we investigated the bird orders in detail, the link between data availability and model performance is present in other orders too (figure 2). Some orders are notoriously difficult to identify to species level, e.g. flies (Diptera) and beetles (Coleoptera), but our models for these perform surprisingly well. The list of species with sufficient observations with pictures for inclusion in the experiment reveals that only relatively easy to recognize species, often with distinct colorations (e.g. ladybugs for beetles) are represented in this subset.

More generally, the requirement that species must have at least 220 citizen science observations with pictures generates a non-random subset of species, and it differs greatly per order how selective this criterion is. Bird species are most frequently reported; 48% of the species present in Norway[22] within the bird orders examined here meet the selection criterion. One of the other orders for which the pattern was found, the dragonflies and damselflies (Odonata), have only 52 species in Norway, of which 44% met the criteria for inclusion. This is in stark contrast to the beetles (1% inclusion), and lichens (2% inclusion), where no clear correlation is found. It is reasonable to assume that for these taxa, the experiment only considers the most recognizable species. If observations were thousandfold, more challenging species could be included, giving a broader range in performances and possibly a similar positive correlation between model performance and data availability.

The consequence of the recognizability bias found here is that as more data is collected, ultimately providing the numbers of pictures needed to train models also on less reported, harder to recognize species, current performance of recognition models cannot be extrapolated to these expanded models. In other words, data that are lacking now are in part lacking because such species are harder to recognize. When such data is added in the future, the performance increase will not be as great as in the past. Besides citizen science, even methods that have no inherent reporting bias, such as automated insect camera traps and trail cameras, can still be subject to recognizability bias. There too, species that are less readily identified will result in more unidentifiable pictures, providing relatively less training data.

Image recognition tools play an important role in maintaining the quality of the large amounts of biodiversity data science and management require. There are

limits to what can be identified from a picture however, and identification tools are needed that rely on more than just pixel information. Models that take into account season, location, sound, etc. can be especially beneficial for difficult species. Still, there is no substitute for the taxonomic knowledge of experts. Preserving this knowledge, and making it available in the form of identification keys is vital. These can be powerful tools to more reliably identify challenging species, in tandem with automatic identification.

Acknowledgements

We are grateful to Rune Sørås, Ingeborg H. Bringslid, and Rienk W. Fokkema for their help in annotating pictures.

Data accessibility

All code is available through Zenodo at `https://doi.org/10.5281/zenodo.6734696`. Bird illustrations in figure 1 are works in the Public Domain made by John Gould (1804-1881), obtained through the Biodiversity Heritage Library[28–30]

Authors' contributions

WK: conception, experimental design, code, analysis, writing. LH: code, text revision. EBN: conception, text revision. RBOH: analysis, text revision. AGF: conception, analysis, text revision.

References

1. IPBES. *Thematic assessment of the sustainable use of wild species of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (eds Fromentin, J.-M. *et al.*) `https://doi.org/10.5281/zenodo.6448567` (2022) (IPBES Secretariat, Bonn, Germany, July 2022).
2. IPCC. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Pörtner, H.-O. *et al.*) (Cambridge University Press. In Press, 2022).
3. Secretariat of the Convention on Biological Diversity. *Global Biodiversity Outlook 5* Sept. 2020.
4. Xu, H. *et al.* Ensuring effective implementation of the post-2020 global biodiversity targets. *Nature Ecology & Evolution* **5,** 411–418. `https://doi.org/10.1038/s41559-020-01375-y` (Jan. 2021).

5. Wetzel, F. T. *et al.* Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in Europe. *Biological Conservation* **221,** 78–85. https://doi.org/10.1016/j.biocon.2017.12.024 (May 2018).
6. Scholes, R. J. *et al.* Toward a Global Biodiversity Observing System. *Science* **321,** 1044–1045. https://doi.org/10.1126/science.1162055 (Aug. 2008).
7. Boakes, E. H. *et al.* Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *PLoS Biology* **8,** e1000385. https://doi.org/10.1371/journal.pbio.1000385 (June 2010).
8. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7.** https://doi.org/10.1038/s41598-017-09084-6 (Aug. 2017).
9. Silvertown, J. A new dawn for citizen science. *Trends in Ecology & Evolution* **24,** 467–471. https://doi.org/10.1016/j.tree.2009.03.017 (Sept. 2009).
10. Hansen, O. L. P. *et al.* Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and Evolution* **10,** 737–747. https://doi.org/10.1002/ece3.5921 (Dec. 2019).
11. Kirkeby, C. *et al.* Advances in automatic identification of flying insects using optical sensors and machine learning. *Scientific Reports* **11.** https://doi.org/10.1038/s41598-021-81005-0 (Jan. 2021).
12. Christin, S., Hervet, É. & Lecomte, N. Applications for deep learning in ecology. *Methods in Ecology and Evolution* **10** (ed Ye, H.) 1632–1644. https://doi.org/10.1111/2041-210x.13256 (July 2019).
13. Weinstein, B. G. A computer vision for animal ecology. *Journal of Animal Ecology* **87** (ed Prugh, L.) 533–545. https://doi.org/10.1111/1365-2656.12780 (Nov. 2017).
14. Wäldchen, J., Rzanny, M., Seeland, M. & Mäder, P. Automated plant species identification—Trends and future directions. *PLOS Computational Biology* **14** (ed Bucksch, A.) e1005993. https://doi.org/10.1371/journal.pcbi.1005993 (Apr. 2018).
15. Ceccaroni, L. *et al.* Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. *Citizen Science: Theory and Practice* **4.** https://doi.org/10.5334/cstp.241 (2019).
16. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org (MIT Press, 2016).

17. Lotfian, M., Ingensand, J. & Brovelli, M. A. The Partnership of Citizen Science and Machine Learning: Benefits, Risks, and Future Challenges for Engagement, Data Collection, and Data Quality. *Sustainability* **13,** 8087. https://doi.org/10.3390/su13148087 (July 2021).

18. Lukhtanov, V. A. Species Delimitation and Analysis of Cryptic Species Diversity in the XXI Century. *Entomological Review* **99,** 463–472. https://doi.org/10.1134/s0013873819040055 (July 2019).

19. The Norwegian Biodiversity Information Centre. *Norwegian Species Observation Service* 2022. https://doi.org/10.15468/zjbzel.

20. GBIF.org. *GBIF Occurrence Download* 2021. https://doi.org/10.15468/DL.TC4W55.

21. Koch, W., Hogeweg, L., Nilsen, E. B. & Finstad, A. G. Maximizing citizen scientists' contribution to automated species recognition. *Scientific Reports* **12.** https://doi.org/10.1038/s41598-022-11257-x (May 2022).

22. Norwegian Biodiversity Information Centre. *Species Nomenclature Database* http://eksport.artsdatabanken.no/Artsnavnebase (2022).

23. Tkachenko, M., Malyuk, M., Holmanyuk, A. & Liubimov, N. *Label Studio: Data labeling software* Open source software available from https://github.com/heartexlabs/label-studio. 2020-2022. https://github.com/heartexlabs/label-studio.

24. ESA. Land Cover CCI Product User Guide Version 2. *Tech. Rep.* https://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf (2017).

25. Sheard, C. *et al.* Ecological drivers of global gradients in avian dispersal inferred from wing morphology. *Nature Communications* **11.** https://doi.org/10.1038/s41467-020-16313-6 (May 2020).

26. Kålås, J. A., Øien, I. J., Stokke, B. & Vang, R. *TOV-E Bird monitoring sampling data. Version 1.6* 2022. https://doi.org/10.15468/6jmw2e.

27. Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830. https://doi.org/10.48550/arXiv.1201.0490 (Nov. 2011).

28. Gould, E., Gould, J. & Lear, E. *The Birds of Europe* https://doi.org/10.5962/bhl.title.65989 (London, Printed by R. and J.E. Taylor, published by the author, 1837).

29. Gould, J. *et al. The Birds of Asia* https://doi.org/10.5962/bhl.title.54727 (London, Printed by Taylor and Francis, published by the author, 1850).

30. Gould, J., Wolf, J., Richter, H. C., Hart, W. M. & Walter of England, Archbishop of Palermo. *The birds of Great Britain* https://doi.org/10.5962/bhl.title.127814 (London, Printed by Taylor and Francis, published by the author, 1873).

Recognizability bias in citizen science photographs

SUPPLEMENTARY MATERIALS

Order	**Species**	**F_1-score**
Passeriformes	*Perisoreus infaustus*	0.901
Passeriformes	*Cinclus cinclus*	0.886
Passeriformes	*Periparus ater*	0.88
Passeriformes	*Bombycilla garrulus*	0.876
Anseriformes	*Aix galericulata*	0.876
Passeriformes	*Certhia familiaris*	0.874
Passeriformes	*Aegithalos caudatus*	0.873
Charadriiformes	*Charadrius morinellus*	0.872
Passeriformes	*Regulus regulus*	0.87
Passeriformes	*Lophophanes cristatus*	0.868
Passeriformes	*Emberiza citrinella*	0.867
Passeriformes	*Garrulus glandarius*	0.865
Passeriformes	*Pyrrhula pyrrhula*	0.863
Passeriformes	*Pinicola enucleator*	0.863
Passeriformes	*Cyanistes caeruleus*	0.86
Passeriformes	*Sitta europaea*	0.856
Passeriformes	*Turdus merula*	0.855
Passeriformes	*Phylloscopus sibilatrix*	0.854
Passeriformes	*Coccothraustes coccothraustes*	0.85
Passeriformes	*Carduelis carduelis*	0.845
Passeriformes	*Motacilla cinerea*	0.839
Passeriformes	*Erithacus rubecula*	0.838
Passeriformes	*Parus major*	0.837
Charadriiformes	*Haematopus ostralegus*	0.837
Passeriformes	*Motacilla alba*	0.837
Passeriformes	*Prunella modularis*	0.836
Passeriformes	*Lanius collurio*	0.835
Charadriiformes	*Phalaropus lobatus*	0.834
Charadriiformes	*Calidris maritima*	0.83

Charadriiformes	*Arenaria interpres*	0.829
Passeriformes	*Phylloscopus inornatus*	0.824
Passeriformes	*Saxicola rubicola*	0.823
Charadriiformes	*Charadrius hiaticula*	0.82
Passeriformes	*Sylvia atricapilla*	0.82
Passeriformes	*Turdus philomelos*	0.819
Charadriiformes	*Gallinago gallinago*	0.819
Passeriformes	*Luscinia svecica*	0.818
Passeriformes	*Plectrophenax nivalis*	0.816
Charadriiformes	*Tringa totanus*	0.815
Passeriformes	*Lanius excubitor*	0.813
Passeriformes	*Turdus viscivorus*	0.812
Passeriformes	*Fringilla coelebs*	0.812
Passeriformes	*Nucifraga caryocatactes*	0.812
Passeriformes	*Passer montanus*	0.808
Passeriformes	*Turdus iliacus*	0.808
Passeriformes	*Emberiza schoeniclus*	0.808
Passeriformes	*Oenanthe oenanthe*	0.807
Passeriformes	*Saxicola rubetra*	0.807
Passeriformes	*Chloris chloris*	0.802
Charadriiformes	*Calidris alpina*	0.8
Passeriformes	*Anthus petrosus*	0.8
Passeriformes	*Motacilla flava*	0.798
Charadriiformes	*Cepphus grylle*	0.795
Passeriformes	*Fringilla montifringilla*	0.794
Passeriformes	*Troglodytes troglodytes*	0.794
Charadriiformes	*Vanellus vanellus*	0.793
Passeriformes	*Carpodacus erythrinus*	0.793
Passeriformes	*Turdus torquatus*	0.793
Passeriformes	*Eremophila alpestris*	0.792
Charadriiformes	*Actitis hypoleucos*	0.788
Charadriiformes	*Pluvialis apricaria*	0.773
Charadriiformes	*Tringa glareola*	0.773
Charadriiformes	*Limosa lapponica*	0.767
Passeriformes	*Ficedula hypoleuca*	0.766
Charadriiformes	*Charadrius dubius*	0.762
Anseriformes	*Cygnus olor*	0.761
Anseriformes	*Mergellus albellus*	0.758
Anseriformes	*Clangula hyemalis*	0.757
Passeriformes	*Muscicapa striata*	0.756

Charadriiformes	*Calidris pugnax*	0.755
Passeriformes	*Phoenicurus ochruros*	0.754
Charadriiformes	*Tringa nebularia*	0.754
Passeriformes	*Acrocephalus schoenobaenus*	0.751
Anseriformes	*Branta leucopsis*	0.747
Passeriformes	*Sturnus vulgaris*	0.746
Charadriiformes	*Limosa limosa*	0.745
Passeriformes	*Calcarius lapponicus*	0.745
Passeriformes	*Phoenicurus phoenicurus*	0.744
Anseriformes	*Mergus merganser*	0.742
Anseriformes	*Bucephala clangula*	0.738
Passeriformes	*Curruca communis*	0.737
Anseriformes	*Tadorna tadorna*	0.734
Passeriformes	*Poecile montanus*	0.732
Anseriformes	*Melanitta fusca*	0.73
Charadriiformes	*Tringa erythropus*	0.729
Anseriformes	*Anas acuta*	0.728
Charadriiformes	*Calidris minuta*	0.727
Passeriformes	*Poecile palustris*	0.725
Passeriformes	*Passer domesticus*	0.723
Charadriiformes	*Calidris temminckii*	0.719
Passeriformes	*Hirundo rustica*	0.718
Charadriiformes	*Numenius arquata*	0.716
Anseriformes	*Mareca penelope*	0.715
Passeriformes	*Corvus frugilegus*	0.714
Anseriformes	*Mergus serrator*	0.71
Passeriformes	*Sylvia curruca*	0.708
Passeriformes	*Turdus pilaris*	0.703
Passeriformes	*Pica pica*	0.702
Charadriiformes	*Uria aalge*	0.699
Charadriiformes	*Chroicocephalus ridibundus*	0.697
Passeriformes	*Hippolais icterina*	0.696
Passeriformes	*Loxia leucoptera*	0.696
Charadriiformes	*Calidris canutus*	0.693
Passeriformes	*Emberiza pusilla*	0.688
Anseriformes	*Cygnus cygnus*	0.683
Passeriformes	*Panurus biarmicus*	0.677
Anseriformes	*Somateria spectabilis*	0.676
Charadriiformes	*Alca torda*	0.675
Charadriiformes	*Rissa tridactyla*	0.674

Anseriformes	*Branta canadensis*	0.671
Anseriformes	*Aythya fuligula*	0.67
Anseriformes	*Anas platyrhynchos*	0.668
Charadriiformes	*Alle alle*	0.663
Charadriiformes	*Calidris alba*	0.652
Passeriformes	*Alauda arvensis*	0.651
Passeriformes	*Corvus monedula*	0.648
Anseriformes	*Anas crecca*	0.644
Passeriformes	*Phylloscopus collybita*	0.643
Charadriiformes	*Pluvialis squatarola*	0.64
Charadriiformes	*Fratercula arctica*	0.64
Anseriformes	*Somateria mollissima*	0.639
Anseriformes	*Anser anser*	0.639
Anseriformes	*Anser indicus*	0.635
Passeriformes	*Acanthis flammea*	0.628
Passeriformes	*Anthus pratensis*	0.627
Charadriiformes	*Sterna hirundo*	0.618
Passeriformes	*Corvus cornix*	0.618
Charadriiformes	*Stercorarius parasiticus*	0.616
Passeriformes	*Phylloscopus trochilus*	0.613
Charadriiformes	*Larus hyperboreus*	0.613
Anseriformes	*Anser brachyrhynchus*	0.608
Anseriformes	*Aythya marila*	0.605
Charadriiformes	*Calidris ferruginea*	0.605
Anseriformes	*Aythya ferina*	0.605
Passeriformes	*Corvus corax*	0.596
Charadriiformes	*Larus fuscus*	0.592
Charadriiformes	*Tringa ochropus*	0.58
Passeriformes	*Locustella naevia*	0.579
Passeriformes	*Carduelis spinus*	0.577
Charadriiformes	*Numenius phaeopus*	0.575
Charadriiformes	*Larus canus*	0.574
Passeriformes	*Carduelis flavirostris*	0.565
Charadriiformes	*Larus glaucoides*	0.564
Charadriiformes	*Larus marinus*	0.559
Passeriformes	*Anthus trivialis*	0.554
Passeriformes	*Regulus ignicapilla*	0.55
Charadriiformes	*Larus argentatus*	0.542
Passeriformes	*Riparia riparia*	0.534
Charadriiformes	*Scolopax rusticola*	0.532

Charadriiformes	*Phalaropus fulicarius*	0.526
Passeriformes	*Sylvia nisoria*	0.523
Anseriformes	*Polysticta stelleri*	0.517
Passeriformes	*Acanthis hornemanni*	0.516
Charadriiformes	*Stercorarius skua*	0.516
Anseriformes	*Melanitta nigra*	0.509
Passeriformes	*Sylvia borin*	0.506
Passeriformes	*Loxia pytyopsittacus*	0.498
Charadriiformes	*Calidris falcinellus*	0.495
Charadriiformes	*Sterna paradisaea*	0.493
Passeriformes	*Lullula arborea*	0.493
Charadriiformes	*Hydrocoloeus minutus*	0.485
Passeriformes	*Pastor roseus*	0.478
Passeriformes	*Loxia curvirostra*	0.477
Passeriformes	*Acanthis cabaret*	0.471
Passeriformes	*Turdus atrogularis*	0.46
Passeriformes	*Ficedula parva*	0.445
Charadriiformes	*Stercorarius longicaudus*	0.442
Passeriformes	*Corvus corone*	0.428
Anseriformes	*Anas clypeata*	0.428
Passeriformes	*Carduelis cannabina*	0.426
Anseriformes	*Anser albifrons*	0.399
Passeriformes	*Acrocephalus dumetorum*	0.397
Charadriiformes	*Calidris melanotos*	0.394
Passeriformes	*Acrocephalus palustris*	0.389
Anseriformes	*Anas strepera*	0.375
Passeriformes	*Delichon urbicum*	0.37
Anseriformes	*Mareca strepera*	0.364
Anseriformes	*Anser fabalis*	0.359
Charadriiformes	*Thalasseus sandvicensis*	0.356
Anseriformes	*Branta bernicla*	0.354
Charadriiformes	*Larus melanocephalus*	0.352
Passeriformes	*Acrocephalus scirpaceus*	0.347
Passeriformes	*Motacilla citreola*	0.318
Passeriformes	*Luscinia luscinia*	0.307
Anseriformes	*Aythya collaris*	0.303
Charadriiformes	*Lymnocryptes minimus*	0.276
Anseriformes	*Spatula clypeata*	0.271
Passeriformes	*Emberiza leucocephalos*	0.244
Charadriiformes	*Larus cachinnans*	0.244

Anseriformes	*Anas querquedula*	0.241
Anseriformes	*Anas carolinensis*	0.215
Anseriformes	*Tadorna ferruginea*	0.215
Anseriformes	*Cygnus columbianus*	0.184
Passeriformes	*Anthus richardi*	0.174
Passeriformes	*Spinus spinus*	0.164
Passeriformes	*Anthus hodgsoni*	0.15
Anseriformes	*Spatula querquedula*	0.149
Passeriformes	*Anthus cervinus*	0.136
Passeriformes	*Linaria cannabina*	0.133
Passeriformes	*Linaria flavirostris*	0.113
Anseriformes	*Anser serrirostris*	0.076

Table S1: Model performances (F_1-scores) for species within the bird orders

Dependent variable	**Parameters**	**Slope**	**Intercept**	$\mathbf{R^2}$	**P-value**
Agaricales F_1-score	Data availability (log)	0.21	0.27	0.15	2.15×10^{-3}
Anseriformes documentation rate	Picture density	0.38	-0.47	0.52	1.51×10^{-7}
Anseriformes F_1-score	Data availability (log)	0.48	-0.97	0.69	4.18×10^{-11}
Anseriformes F_1-score	Documentation rate	-1.52	0.63	0.19	4.64×10^{-3}
Anseriformes F_1-score	Picture density	-1.02	1.95	0.31	2.10×10^{-4}
Asparagales F_1-score	Data availability (log)	0.31	0.01	0.34	0.0117
Asterales F_1-score	Data availability (log)	0.59	-0.71	0.51	5.92×10^{-4}
Charadriiformes documentation rate	Picture density	0.34	-0.43	0.76	5.51×10^{-18}
Charadriiformes F_1-score	Data availability (log)	0.32	-0.34	0.4	2.49×10^{-7}
Charadriiformes F_1-score	Documentation rate	-0.9	0.7	0.28	3.45×10^{-5}

Charadriiformes F_1-score	Picture density	-0.42	1.25	0.39	3.52×10^{-7}
Coleoptera F_1-score	Data availability (log)	0.13	0.57	0.06	0.273
Diptera F_1-score	Data availability (log)	0.53	-0.63	0.19	0.0326
Lecanorales F_1-score	Data availability (log)	0.2	0.31	0.12	0.118
Lepidoptera F_1-score	Data availability (log)	0.14	0.49	0.05	9.50×10^{-5}
Odonata F_1-score	Data availability (log)	0.46	-0.53	0.49	2.02×10^{-4}
Passeriformes documentation rate	Picture density	0.3	-0.36	0.55	1.59×10^{-19}
Passeriformes F_1-score	Data availability (log)	0.54	-1	0.63	4.31×10^{-24}
Passeriformes F_1-score	Documentation rate	-0.85	0.71	0.11	5.19×10^{-4}
Passeriformes F_1-score	Picture density	-0.5	1.37	0.23	1.68×10^{-7}
Passeriformes picture quality	Habitat openness	-0.12	5.93	0.26	5.53×10^{-8}
Polyporales F_1-score	Data availability (log)	0.31	-0.11	0.28	2.18×10^{-3}

Table S2: Metrics collected for species within the bird orders

Paper IV

Clavis: an open and versatile identification key format

Wouter Koch[1,2] Hallvard Elven[3] Anders G. Finstad[1]

1: Department of Natural History, Norwegian University of Science and Technology, Trondheim, Norway; 2: Norwegian Biodiversity Information Centre, Trondheim, Norway 3: Natural History Museum, University of Oslo, Oslo, Norway

Preprint on bioRχiv: doi:10.1101/2022.05.26.493630

Abstract

The skills and knowledge needed to recognize and classify taxa is becoming increasingly scarce in the scientific community. At the same time, it is clear that these skills are strongly needed in biodiversity monitoring for management and conservation, especially when carried out by citizen scientists. Formalizing the required knowledge in the form of digital identification keys is one way of making such knowledge more available for professional and amateur observers of biodiversity. In this paper we describe Clavis, a modern open format for capturing knowledge required for taxon identification through digital keys, allowing for a level of detail beyond that of any current key format. We exemplify each concept using Pokémon as a fictional taxonomic group.

Introduction

Distinguishing biological taxa from one another is a necessity in biodiversity monitoring. As species are going extinct at an unprecedented rate[1,2], we need to monitor as much of nature as we can to identify population statuses and trends. Meanwhile, research and management is currently facing a "taxonomic impediment", where taxonomic knowledge is gradually disappearing from the scientific community[3].

Paradoxically, while taxonomic expertise is becoming scarce, species observations are being reported like never before. The bulk of the observational data currently available originates from citizen science, however, in which observations are made by non-professional volunteers[4]. There are large taxonomic biases in the data collected from these sources. For example, of publicly available data from

the Global Biodiversity Information Facility (the largest aggregator of biodiversity observations in the world[5]), some 67% of over 2 billion observations pertain to birds[6]. This means that less charismatic species, and species that are more difficult to identify, generally remain severely under-reported[7]. Many such species play important ecological roles or can serve as indicators of broader ecosystem health, but the fact that few observers are able to reliably identify these species leads to their underreporting, thus obscuring important trends from the view of science and management[8].

The knowledge needed to identify species resides in large part in the heads of taxonomic experts. This invaluable experience is slowly disappearing as newly educated biologists are, to an increasing extent, trained with focus on skills in genetics, data management, and biodiversity informatics rather than traditional taxonomy. This training, and an academic reality of shorter periods with funding and more temporary contracts, gives fewer researchers the opportunity to invest the years of experience needed to become truly knowledgeable with regards to any substantial taxon, as was more commonplace in the past[9].

Identification keys are one of the most important tools used by taxonomists to identify taxa, as well as to share with others the information needed for distinguishing taxa. No taxonomist studying a large taxon in depth can get around identification keys, and keys are often the starting point when trying to learn a new organism group. Identification keys can, however, be quite challenging to use, especially for novel users, and their use often represents a significant barrier for the prospective learner. For this and other reasons, the use of identification keys is receiving less focus in most current biology curricula. Furthermore, as taxonomic knowledge is becoming increasingly fragmented, existing identification keys in literature are gradually becoming outdated as taxonomies are changing according to new insights.

These issues are exacerbated in the context of citizen science. Whereas prospective taxonomists have a professional incentive to learn to use a key, a steep learning curve can quickly put off the more casual citizen scientist. Citizen science, with its unrivaled amounts of data, but also added concerns regarding the reliability of the taxon identifications[10,11] and the pronounced over-representation of more charismatic taxa[12–14], stands especially much to gain from good identification keys with a low user threshold.

Digital identification keys address these challenges in a number of ways. The formalized and machine-readable way in which the data are stored allows for easier revision, as well as multiple options for their display to the end user, tailored to their level of experience. Digital keys allow for the representation of more complex relationships between species' characteristics and identifications inherent to biological complexity. These relationships cannot as easily be represented in a linear form as is used in traditional, paper-based keys. With a suitable interface, digital keys open up the cumulative experience from taxonomic experts to a broader public like citizen scientists, while potentially being more reliable and educational

than other identification methods such as automated image recognition. In contrast to paper-based keys, digital keys can also be combined with one another, integrating the information from several separate keys into one seamless user experience. And conversely, a digital key can easily be limited to just a subset of the taxa in it, for instance only taxa belonging to a certain habitat or geographic range. Another benefit is the possibility of saving the user input and the key together with the observational data, as a means of transparent identification that can be reviewed at any time in the future for quality control or in the case of new taxonomical insights.

Having a fully open and well-defined, platform-independent format is essential in ensuring that identification keys and the tools needed to display and create them remain both interoperable, interpretable across platforms, and freely available to all. A number of digital identification key formats exist[15,16], but these come with a number of limitations in what they can represent, their ease of use, openness, etc. A well-defined format addressing this alleviates the technical burden in capturing taxonomic knowledge for future use.

In this manuscript we document Clavis; an open format for identification keys, aiming to cover the aforementioned requirements in an open and lightweight manner, serving as a way to store and exchange crucial taxonomic knowledge. Clavis is intended to capture all the requirements of traditional keys while adding the flexibility and complexity possible with digital keys. This means that the format is well suited for representing most, if not all, existing keys, both paper-based and digital, as well as for designing new keys. The name Clavis means "key" in Latin, and is a recursive acronym for Clavis Lightweight And Versatile Identification Schema. This article describes the Clavis format itself, example code handling the business logic of keys described using Clavis will be published separately at a later time. Descriptions of each of Clavis' data types and emerging properties are illustrated here using fictional taxa that are discretely defined and not subject to taxonomic debate or change, while exhibiting the required complexity for a demonstration of the more intricate features of the format. Our fictional creatures of choice for these examples are a selection of Pokémon[17].

Material and methods

Identification keys exist in many forms, but all are built on the same basic principle. The user is asked a number of questions about the entity being identified, and by answering these questions, taxa are excluded from consideration until (ideally) only one taxon remains, which is the result of the identification process.

Traditional single access keys consist of a decision tree with a single fixed path from beginning to end for each possible outcome. On choosing an answer from the alternatives for each of the questions, the user is directed to the next question to be answered, until they end up with a result taxon. Such keys can be either

dichotomous, meaning that each question always has only two alternatives, leading to a bifurcating decision tree, or they can be multichotomous, meaning that each question may have more than two alternatives. One upside of single access keys is that they are easy to represent on paper. A drawback is that a user has to go through all questions in a specific order, having one and only one question to answer at any given time. This means that one cannot use only part of the key for any subset of taxa, nor easily go back and redo a question. Also, if a question cannot be answered, there is no way to proceed.

An alternative approach is the multiple access key. Such keys are typically stored as matrices where taxa and questions (characters) are stored as rows and columns in a tabular format, with cell values linking the taxa to their characters. Each character may have two or more possible answers (states). In a fully populated matrix key, every character is assigned a value (i.e. is scored) for every taxon. The user then has access to all the questions/characters at once and is not required to answer them in any particular order. Choosing any alternative for any of the questions will exclude from consideration all taxa that are not compatible with that alternative, bringing the user a step closer to the answer. The upside of this approach is that the user has several paths to the answer, and can choose to avoid questions that are difficult or impossible to answer. The downside is that the number of choices can easily be overwhelming, and it is in large part left to the user to try to choose the best path. Also, in real life situations it is very rare to find a set of characters that are both possible to score in a meaningful way for all the taxa in the key, and sufficient for distinguishing between them. More often, only some characters will be possible to score for all taxa, whereas others may be essential for distinguishing certain taxa in the key, yet be inapplicable to others.

To address this, a matrix can be filled out sparsely instead, meaning that not all characters are scored for all taxa. The interface can then display only those characters that are in fact scored for all the taxa currently under consideration. As taxa are being excluded by the user's choices, further characters that are scored for all the remaining taxa will become available for answering. With this approach, it is possible to make a matrix key that behaves equivalently to a single access key. But the key can also be made to contain several distinguishing characters for any set of taxa, so that the user has more than one choice at each step and is thus not restricted to having to answer only one question at a time in a specific order.

The sparse matrix approach is more powerful than both the single access and the full matrix approach, in that both of these approaches are subsets of what the sparse matrix can represent. An important advantage of either matrix approach over the single access key is that the key can easily be restricted to only a subset of the taxa. This means that the user will not need to traverse the whole key if they wish to exclude one or more taxa a priori. Still, there is much additional information one would ideally like to store about taxa, characters, and the relationships between the two, which cannot be easily represented in a tabular format. For instance, the tabular format is not well suited for dealing

with polymorphism in a straightforward way, i.e. the situation where a taxon can have more than one possible value for a given character. Also, it may often be desirable to treat taxa hierarchically rather than as a simple list of equally ranked entities, something which is difficult to do in the matrix format. Thus, representing taxonomic knowledge in this tabular format restricts the complexity of the information that can be stored.

The method of storing and exchanging taxonomic knowledge described here, is developed as a non-tabular multiple access key. The core of any Clavis key is a collection of statements. A statement links a taxon to a character, and specifies a state or numerical value the taxon has for that character (see Fig. 1). E.g. the color (character) of species x (taxon) is red (state), or the length in millimeters (character) of genus x (taxon) is 4-11 (numerical range).

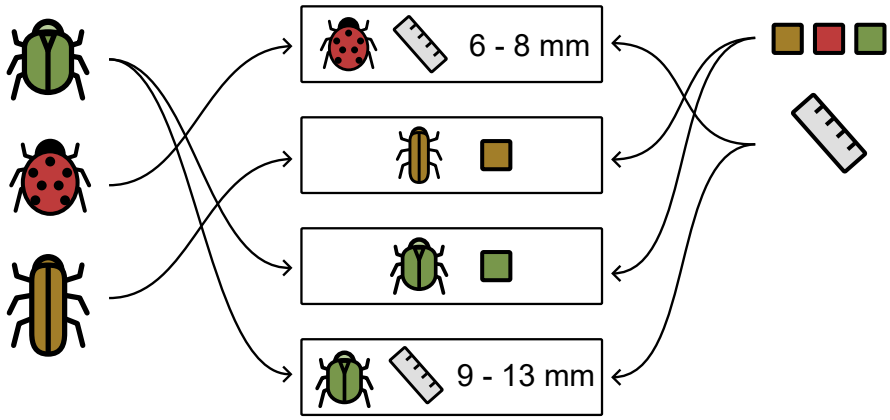


Figure 1: **The core concepts of any Clavis key.** Taxa (left) and characters (right) are connected through statements (center). Each statement refers to one taxon and one character, linking the two by a value, usually one of the possible states of the character, or a numerical value or range.

Additional information can be linked to taxa, characters, states, and statements in a number of ways, capturing knowledge and relationships that cannot be represented in a tabular form in a practical way. Such information includes media elements, textual descriptions, and contextual information such as geographic scope. The Clavis format incorporates all the possibilities of the previously described approaches, so that any single access or matrix key can be transcribed to it.

Clavis compliant keys are written in JSON (JavaScript Object Notation)[18], which has become the *de facto* standard for data exchange over the internet, as it is an open, flexible format with widespread use and support within all modern programming languages. Clavis itself is a JSON-schema[19], a formal definition of

the structure and content of a valid Clavis JSON file. JSON-schemas are machine readable and can be used to automatically validate the compliance of a JSON file in modern code editors, highlighting any issues that need to be solved. While it is possible to manually write the JSON of a Clavis-compliant key, it is generally not practical. In order to facilitate key creation and editing, one would generally provide taxonomists with a key editing interface that lets them easily record their knowledge, and let the key editor generate the corresponding JSON.

Implementation

As an exchange and storage format, Clavis does not dictate how it is to be implemented. Different interfaces can implement it differently, depending on the purpose of the interface and the intended user group. To this end, interfaces serving to edit or display keys may disregard certain non-essential functionality supported by Clavis. One could for example create a key editor not supporting media files or multilingualism. Other aspects of the format are crucial however, and require support and unambiguous interpretation.

Identification is a matter of excluding taxa, and is done by letting the user select a state or numerical value for characters. Each time the user provides a new fact in this way, all taxa with conflicting statements are excluded (see Fig. 2). The user can also be given the option to exclude a state rather than affirm one if there are more than two states to choose from.

Whenever taxa are excluded, characters that were previously hidden may now have become relevant if all the remaining taxa have a statement pertaining to it, as illustrated in Fig. 2. These characters should then be made visible to the user. The act of excluding taxa will also affect which of the possible states of a given character are relevant, even if it has not been answered directly by the user. If all the remaining taxa are known not to have a certain state, that state can be disabled. In such cases, the state still needs to be visible to the user as it provides context to the remaining alternatives, but it should not be possible for the user to select it. Characters that have not been answered, but that only have a single possible answer for all remaining taxa, can be hidden. In all cases, only characters that are linked by statements to all the currently non-excluded taxa should be shown to the user. Clavis also allows for the inclusion of explicit dependencies, where the relevance of a character is determined by the answers given on one or more other characters.

Things become more complicated when allowing the user to undo previous answers. Undoing an answer might render a character that was subsequently answered irrelevant again. To avoid answers to this now irrelevant character from eliminating taxa as possibilities, an implementation needs to be able to remove or ignore answers on characters that have been rendered irrelevant.

The provided schema ensures that a key adheres to the Clavis format in a technical sense, but it does not ensure that any compliant key is logically complete

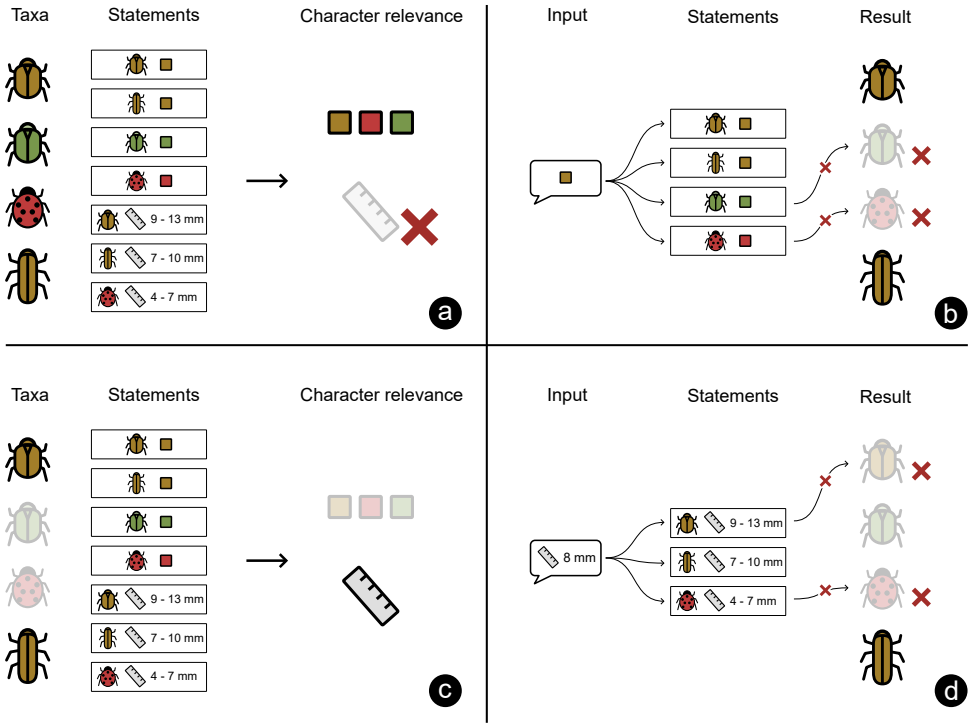


Figure 2: **The identification of a specimen.** When considering all taxa, the character about size is not relevant, as one of the taxa has no statement for size, so only the character about colors is relevant (a). The user gives an input, which is compared to the related statements, upon which 2 taxa are eliminated (b). Considering the now remaining taxa, the character about size has become relevant (c). The user gives another input, and based on the related statements, one additional taxon is eliminated, leaving one result (d).

and consistent. One can make a valid key that contradicts itself or that does not contain sufficient information to reliably distinguish between certain taxa. It is advisable to implement checks for this in any key-editor, and when keys by third parties are to be displayed, to also ensure such cases are handled by the end-user interface.

To illustrate the features supported by the Clavis format, we here describe features and give some example uses. The examples relate to fictional creatures that are not subject to taxonomic debate or change. For this we have chosen Pokémon as they appear in the mobile game Pokémon Go, as these are clearly defined, yet exhibit the complexity required to demonstrate the various features of the format. Keys for natural taxa would rarely use as much of the possible functionality as we demonstrate here in a single key. The key is valid in accordance with the JSON-schema, but it does not contain all the information needed to distinguish between all of the Pokémon mentioned in it. We also provide an

example of a non-artificial key that does identify all taxa in it, but that does not aim to demonstrate all possible features that Clavis supports. This key covers all the species of titmice in Norway and can be found in the S3 Appendix.

Results

The main components of the Clavis JSON-schema are the taxa that the key is designed to distinguish between, the characters describing relevant properties, and statements connecting taxa to characters through states or numerical values. Additionally, the schema defines a number of metadata fields, as well as custom data types, that can be referred to in various places in the key, such as when referring to a person as a creator of a picture or linking a picture to a taxon.

Format overview

An overview of the elements in an identification key defined by Clavis. Elements with an asterisk (*) are mandatory.

Key metadata

Title*	The name of the key
Schema*	The url of the version of Clavis that the key adheres to
Media	A media element for the key, such as a logo or icon
Description	A short, extended, and/or external description of the key
Audience	A description of the intended audience
Source	Name and/or link to the source the key is based upon
Geography	Polygon and/or name of the region where the key is valid
Roles	Primary contact, creators*, contributors, publishers of the key, as references to persons and/or organizations
License*	An url to the license text the key is licensed under
Language*	The language(s) the key supports
Dates*	The dates the key was created and last modified (the version)
Identifier*	An id for the key, that remains stable over versions
Url	Where the key is hosted, so that new versions can be retrieved

Key content

Taxa*	A flat or hierarchical list of taxa the key is able to distinguish between. The goal of the interface is to eliminate all but one taxon.
Characters*	A list of characters used to distinguish between taxa. These are the questions that are presented to the user. A character can be categorical or numerical. When categorical, it has a list of states that are the relevant alternatives for this character.
Statements*	Elements connecting a taxon to a character through a state or numeric value. This is the core knowledge captured by the key: which taxa have which states or numerical values for which characters, thus distinguishing them from one another.

Data types

Person	The name*, contacts, media elements, and affiliations of a person.
Organization	The name*, contacts, and media elements of an organization.
Taxon	The scientific name, author, vernacular name, label, media elements, rank of a taxon. Info on whether it serves as an end-point or not. It can have a set of children, which are the underlying taxa. An external reference can define where info is to be retrieved from. It may have a geographic distribution as an object or external service, to assess where it occurs. A follow-up key as a url or external service reference can provide info on where a more detailed identification can be done.

Character The name* and states describing a property a taxon can have. It can have media elements and descriptions clarifying the character, and a user requirement. A logical premise can specify what other user input has to be given before the character may be presented. A character can be of the types "exclusive" (default, multiple choice where options exclude one another), "non-exclusive" (multiple choice with multiple answers possible), and "numerical" (the answer is a number). A character has states defining the possible answers if it is not numerical, and a min, max, unit and step size if it is numerical.

State A possible answer for a non-numerical character. Has a title and can have multiple media elements and descriptions clarifying the state.

Statement A connection between a taxon and a character through a value (either a state or a numerical range). It defines how frequently and in which context the taxon has this value for that character. A statement can have any frequency from 0 to 1, to indicate that the taxon always (1), never (0) or in some cases (values between 0 and 1) has this value for the character. A numerical value can be a single value or a range. Statements can contain references to a geographic distribution (or a service providing one), defining where this statement is valid. Media elements and descriptions can be added describing the relationship between the taxon and character in more detail.

User requirement A user requirement can have a title, media elements and descriptions describing certain skills, equipment or other requirements needed to evaluate a character. It can also have a warning text to alert the user of these requirements. It can be used to guide a user where necessary, or help the user decide whether to skip more challenging characters.

External service A reference to an external service and its documentation. The creator of an interface can then choose to implement this external service so that e.g. taxon names are retrieved from an up to date repository by the provided stable identifier.

Media element A media element contains the data needed to display multimedia files. It can refer to different versions for different languages, and can have different versions for different media dimensions. It contains the required metadata such as width and height (images and video), length (sound and video), as well as creators, contributors, publishers and a license. Files can be urls to where the correct version of the file is to be found, or directly contain a base64 or svg encoded file. It supports external services to retrieve data from elsewhere.

Statements

The core element of any key is the statement. It defines a property of a taxon, separating it from other taxa that have conflicting statements. In JSON code, a single statement takes the following form, in this case stating that Pikachu is an electric type Pokémon.

```
{
   "id": "statement:pikachu_is_electric",
   "taxon": "taxon:pokemon_025",
   "character": "character:type_of_pokemon",
   "value": "state:pokemon_type_electric",
   "frequency": 1
}
```

While all core concepts are described with examples, not every combination of concepts are exemplified here. So while both multilingualism and descriptions are demonstrated, there is no example of multilingual descriptions. Such features do follow the same logic as the examples provided, and are all specified in the JSON schema.

Key metadata

Very few parameters are required on the top level of the key. Apart from the content of the key (taxa, characters, and statements linking the two), a key is expected to refer to the version of the schema with which it complies, a title, language, license, a creator, the date at which it was last modified and an identifier that is to be kept stable across versions. As the creator has to be a reference to a person entity, at least one person needs to be defined as well.

Example*: See lines 2 - 10 in the S2 Appendix for the corresponding JSON.*

Taxa

Eliminating all but one taxon is the goal of the user, and taxa are the units that all characteristics are connected to. Taxa can be provided as a flat list, but they can also be structured hierarchically, commonly, but not necessarily, adhering to their phylogeny. Statements can in such a hierarchy be connected to higher taxa, reducing a lot of the repetition of traits shared within a taxon that one would get when using a flat list of taxa. If a statement is tied to a higher taxon in the hierarchy, it is implied that all the underlying taxa share the same statement.

In addition to biological taxonomic units, one can also define sub-groups within a taxon. Examples can be different sexes, morphs, or species complexes. Contrary to regular taxa, such subdivisions of taxa are not standalone taxonomic units. As such, they do not have their own scientific name but rather a label that adds specificity to their parent taxon. For example, the label of a sub-group specifying the sex of Pikachu would simply be "♀" rather than "Pikachu ♀" as it already relates to the parent taxon "Pikachu". For a default form of a taxon, the label can be an empty string.

The goal of the key is to provide a way to eliminate all but one taxon; the result of the identification process. The key should stop asking questions once the user has narrowed down the possible outcomes to a single endpoint. In a flat list of taxa, every taxon in the list will be an endpoint. In a hierarchical list, the lowermost taxa in the hierarchy will be the endpoints by default. This can be overridden, however, by explicitly tagging a taxon higher up the tree as an endpoint. In this case, information on lower taxa may be displayed if identified while using the key, but no additional questions are asked once the endpoint has been determined. This can be used, for instance, to display an image of the relevant sub-group for a taxon, so that the taxon images reflect the input from the user.

***Example**: Clavicula, Clavis and Clavissima are all species within the Clavidae. They each have a default and a "Shiny" morph, and in this key, the morph is the endpoint. The default Clavissima morph also has a subdivision into the sexes, but as the morph is tagged as an endpoint, the user will not be asked further questions to determine the sex once the morph has been determined. See lines 25 - 94 in the S2 Appendix for the corresponding JSON, and Fig. 3 for a graphical representation.*

Characters, statements and frequencies

One of the possible values for a statement (the relationship between a taxon and a character) is the id of a state. A statement also may have a frequency: the proportion of cases where individuals of this taxon have this state for this character. It can be set to 0 or 1 or any value in between.

In the default setting, the states of a character are interpreted as being mutually exclusive. If there is a character with states "blue" and "red", a specimen may be either blue or red, but not both at the same time. This means that a taxon that is

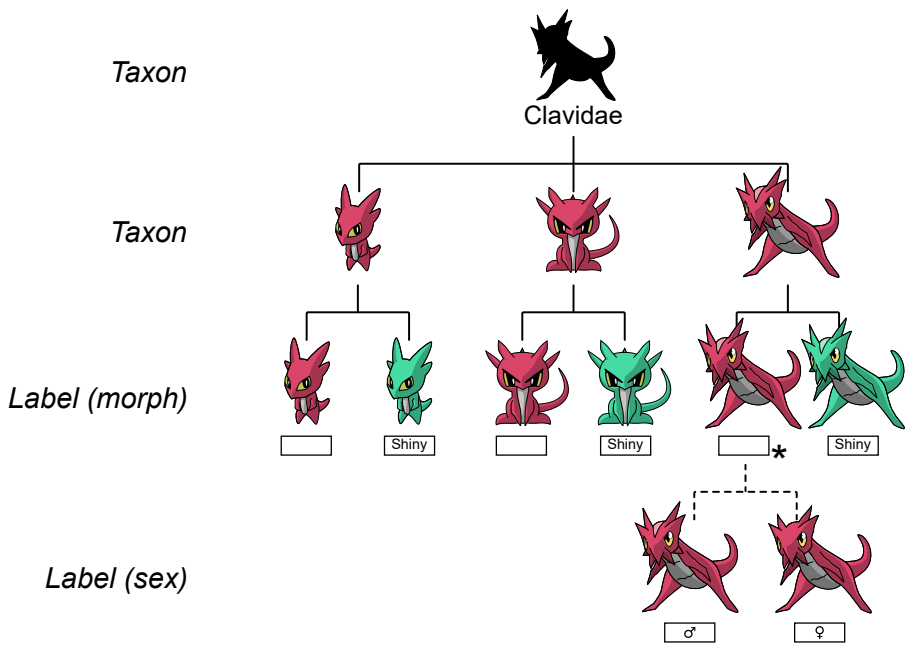


Figure 3: ***A taxon tree of the Clavidae-line****: Taxa tagged as endpoints are indicated with an asterisk (*). This key continues until the correct morph ("Shiny" or default) is known. As the morph of Clavissima (the species to the right) is defined as an endpoint, no questions regarding the sexes of Clavissima will be asked once the correct species and its morph have been determined.*

noted as always being blue is known never to be red, and vice versa. This default behavior can be overridden, however, by defining the character as non-exclusive. In this case, stating that a specimen is blue does not exclude the possibility that it is also red.

Example*: Pikachu is always an electric type. See lines 261 - 267 in the S2 Appendix for the corresponding JSON.*

Example*: Pikachu is never blue. See lines 268 - 274 in the S2 Appendix for the corresponding JSON.*

Example*: Pikachu has a double-lobed tail in 50% of the cases. See lines 275 - 281 in the S2 Appendix for the corresponding JSON.*

Example*: Pikachu contains the colors yellow, black and red. The closely related Raichu contains yellow, brown, white and red. Setting the color character to be of non-exclusive type allows the user to indicate all the colors that each of them has. See lines 177 - 203 in the S2 Appendix for the corresponding JSON, and Fig. 4 for a graphical representation using Clavidae.*

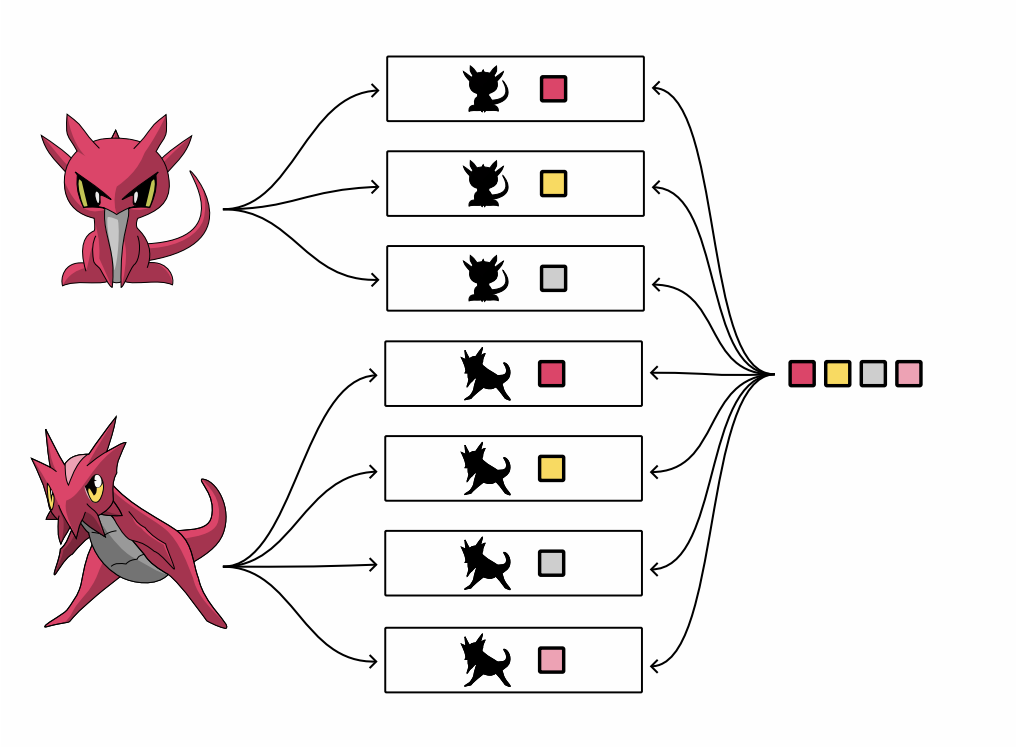


Figure 4: ***The colors of Clavis and Clavissima****: Clavis and Clavissima both have several colors. In its default setting, the color character will allow the user to select only one color, which will imply that the species does not have the other colors. By defining the character as non-exclusive, the user can pick and/or exclude freely from the list of possible colors.*

Multilingualism

Rather than specifying the language of the key by referring to a single ISO 639-1 code, one can make a key multilingual by specifying an array of language codes that are supported. When doing so, strings within the key that differ between languages have to be given as localizedStrings; objects containing a version of each of the supported languages. To support different script types, also strings like people's names support localizedStrings.

Not only strings can have different versions for different languages. Links to external online resources can refer to separate language versions (localizedUrls), and images (e.g. containing text) can have different versions for different languages too (localizedMediaElements).

Example*: A creator name requiring different transcriptions within a multilingual English/Ukrainian key*

```
"language": ["en", "uk"],
"creator": "person:wouterkoch",
```

```
"persons": [
   {
     "id": "person:wouterkoch",
     "name": {
       "en": "Wouter Koch",
       "uk": "Вауrер Kox"
     }
   }
]
```

***Example**: A creator name that needs no translation, and a character that does, within a multilingual English/Dutch key*

```
"language": ["en", "nl"],
"creator": "person:wouterkoch",
"persons": [
   {
     "id": "person:wouterkoch",
     "name": "Wouter Koch"
   }
],
"characters": [
   {
     "id": "character:eye_color",
     "title": {
        "en": "Color of the eyes",
        "nl": "Kleur van de ogen"
     },
     "states": [
       {
         "id": "state:blue_eyes",
         "title": {"en": "Blue", "nl": "Blauw"}
       },
       {
         "id": "state:red_eyes",
         "title": {"en": "Red", "nl": "Rood"}
       }
     ]
   }
]
```

***Example**: An organization with names and urls within a multilingual Norwegian/English key*

```
"language": ["no", "en"],
"publisher": "organization:ntnu",
```

```
"organizations": [
   {
     "id": "organization:ntnu",
     "name": {
       "no": "Norges teknisk-naturvitenskapelige universitet",
       "en": "Norwegian University of Science and Technology"
     },
     "url": {
       "no": "https://www.ntnu.no",
       "en": "https://www.ntnu.edu"
     }
   }
]
```

Persons and organizations

Persons and organizations can have multiple roles in different contexts. A person can be a contact person for a key, and one or more persons can be creators or contributors to a key. Persons can also be creators or contributors to media files. A person can have one or more organizations as their affiliation, and organizations can have a person as a primary contact. An organization can be the publisher of a key or media file, and the primary contact for a key can also be an organization instead of a person. Persons and institutions can have resources such as urls and mediaFiles connected to them, e.g. institutional websites, portraits and logos.

Example*: Definition of a person and an organization. See lines 11 - 23 in the S2 Appendix for the corresponding JSON.*

Geographic and taxonomic scope

Geographic regions can be defined using a name and/or GeoJSON MultiPolygon. On the top level, a "geography" field can be specified to indicate what geographic region the key covers. One can also define a geography at the taxon or statement level, to provide information on where the taxon occurs, and in which region a taxon has that particular relationship to a character. Within a region specified as the geography of a statement, the statement with a geography takes precedence over the conflicting statement without a stated geography.

Example*: Kangaskhan only occurs in Australia. See lines 95 - 146 in the S2 Appendix for the corresponding JSON.*

Example*: The rain form of Castform has a higher frequency in the notoriously rainy Norwegian city of Bergen. See lines 299 - 344 in the S2 Appendix for the corresponding JSON.*

Logical premises

Not all characters are meaningful in every context, and it is sometimes desirable to be able to specify conditions that must be met before a character is shown to the user. Such a condition may be that the user has given a certain answer to some other character in the key first. To this end, characters can have a logical premise, specifying which facts have to be established for it to be relevant. Logical premises can also relate several facts combined, including numerical values, and consist of strings using unary and binary operators in JavaScript notation.

Logical premises are seldomly required to make a key work, but can be used to ease the identification process. If characters are scored only for those taxa that they are relevant for distinguishing, this in itself will normally ensure that such characters stay hidden from the user until they are relevant to show. However, a situation may arise in which some of the taxa in the key are polymorphic, i.e. if they either may or may not have a certain property. In this situation, a character referring to this property would normally become visible to the user because all the remaining taxa *can* have it, which will likely cause confusion if the actual specimen being identified does *not* have it, or if it is not observable. By introducing a logical premise, such a character may be hidden until the user has explicitly indicated that the specimen does in fact have the property.

***Example**: Pikachu caught during special events have hats, in contrast to those caught outside of such events. Other Pokémon (e.g. Honchkrow) always have hats. This allows for the characteristics of the hat to be used in identification, but not before it has been established that one is not only left solely with Pokémon that can have hats, but when it has been established that the target Pokémon in fact is wearing one. See lines 220 - 248 in the S2 Appendix for the corresponding JSON.*

Numerical values

The possible values that a character can have can take the form of a set of discrete states, such as “present”/“absent”, or “red”/“blue”. They may, however, also come in the form of numerical values, such as counts or measurements. In a key this can be implemented with a character of the type “numerical”. Numerical characters specify a minimum and maximum that the value can have, a step size and its unit. Statements specifying values for numerical characters to taxa can define a range, or a single value. If more advanced metrics are needed for the key, such as a probability function over the numerical values for use in Bayesian analysis, this can be implemented through a call to an external service (see below).

***Example**: The weight of a Pokémon varies, both between individuals within a species and between species. To distinguish Pokémon based on their weight, this is added as a numerical character, where weights can be specified in whole kilograms within a given range. A statement can then be included to register the possible weight range of a given taxon. See lines 249 - 258 (character) and lines 289 - 298*

(statement) in the S2 Appendix for the corresponding JSON.

External services

Various units in a key may be subject to changes that are ideally managed outside the key, in designated centralized systems. Taxon names can change, as can urls to supplemental information, media files, geographic ranges etc. Other parameters, such as the probability distribution of a numerical character for a given taxon, may be too detailed for direct inclusion in a key. It is best practice to not duplicate such resources, but to harvest these via an Application Programming Interface (API) or other service interface. To facilitate this, Clavis allows the specification of external services, where documentation for use of the service can be linked. Various units in the key, such as taxa, can refer to such external services through one or more externalResources, connecting the service and the relevant external id to the taxon at hand.

External services should not contain information critical to the workings of the key, as not every implementation can be expected to contain the necessary code to retrieve the information the service provides. External services are useful, however, for features steering presentation, and can provide complex information that depends on various user inputs and other contextual information. To sort taxa by probability, for example, an external service may provide probability scores for taxa based on the geographic location of the user, the season, and properties like coloration and size of the specimen. An interface that does not implement calling this service will simply not sort the taxa by probability, but will still be fully functional.

***Example**: To allow for features such as the retrieval of the name of a Pokémon in different languages, one can use a Wikidata query. See lines 432 - 436 (service) and lines 47 - 50 (externalResource) in the S2 Appendix for the corresponding JSON.*

***Example**: A hypothetical API returns the probability for a provided taxon, given its location and size. See lines 437 - 443 in the S2 Appendix for the corresponding JSON.*

Required expertise

Some characters are harder to evaluate than others, and may even require special equipment. To warn and assist users, a key can contain userRequirements, describing such required skills or tools. These requirements can be connected to characters so that the user might filter out characters requiring skills or equipment they do not have, be presented with additional info to complete the task, or simply be warned.

***Example**: the weight of a Pokémon can be a useful characteristic as their weights vary between types. One can only weigh a Pokémon once it has been*

caught, however. The user needs to be aware of this so that in situations where a Pokémon cannot be caught (e.g. when identifying from a picture or being out of Pokéballs) such questions can be ignored. One can also add a guide on how to catch a Pokémon in order to weigh it. See lines 249 - 258 (character) and lines 417 - 429 (userRequirement) in the S2 Appendix for the corresponding JSON.

Media elements

Illustrations greatly improve the usability and aesthetics of any key. Most entities in a Clavis key can contain a reference to an image; taxa, characters and states, but also persons, organizations, userRequirements, statements, etc. To facilitate this, Clavis defines mediaElements, containing one or more mediaFiles. These mediaFiles can contain and/or link to image, sound, or video files, using an array of mediaFiles to allow different sizes of the same file to be included.

Example*: The states in the tail character of Pikachu refer to images illustrating the different tail shapes. See lines 207 - 218 (states) and lines 446 - 496 (mediaElements) in the S2 Appendix for the corresponding JSON.*

Descriptions

There are many aspects of an identification key where a more elaborate description is desirable or even required. To this end, many elements can have a short and/or extended description in valid markdown notation, as well as a url to an online description.

Example*: A description can be used to explain to the user how to catch a Pokémon where relevant, by including a short description and a link to a page with more information. See lines 422 - 429 in the S2 Appendix for the corresponding JSON.*

Followup keys

Once a key has identified a taxon as far as it is intended to, i.e. when it has narrowed it down to an endpoint taxon, the user is presented with the result. It may however be possible to further determine the result. If a key exists somewhere that can help with this, it can be referred to as a followup key from the relevant taxon. It can be either a url, or a reference to an (external) service. This feature can also be used to split keys into several smaller keys that refer to one another through this mechanism.

Example*: Once the Pokémon is determined to be a Pikachu, the user can be advised that there is a key to determine which of its many possible costumes the Pikachu is wearing. See line 56 in the S2 Appendix for the corresponding JSON.*

Discussion

The examples provided here illustrate the versatility of Clavis as a key format. Several of its features are, to our knowledge, not supported by any other format, nor is the totality of its features.

A crucial aspect in identifying any taxon is the geographic origin of the target specimen. Primarily, it dictates which taxa are candidates for its identity, and thus which key(s) can be used for it and which taxa within the keys are to be considered. Secondarily, it dictates the possible traits of the taxa, insofar as these vary geographically. By referring to external services for geographical information, keys can be made to directly benefit from Species Distribution Models hosted elsewhere, ever improving as more data and improved methodology become available.

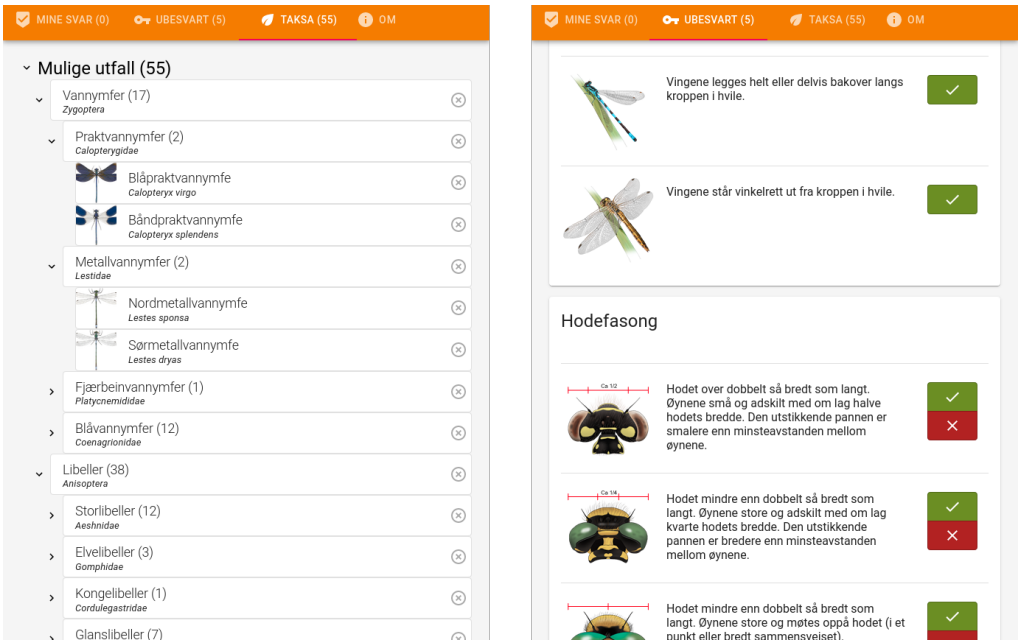


Figure 5: **An example implementation of a Clavis key graphical user interface.** Based on a working Odonata key previously published by the Norwegian Biodiversity Information Centre. A hierarchical list of taxa (left) is reduced by providing input on characters (right). Only characters relevant for all remaining taxa are shown.

It is important to realize that keys are designed to function as a whole, and that its contents need to be regarded in context of the key. Characteristics defined and scored in the context of distinguishing between taxa within a limited taxonomic and geographical scope can be misleading when put in contexts other than that of the key. It may in some cases be possible to extract traits from a key for use in a different context, such as for a taxon diagnosis or a trait database. Taxonomic

knowledge is required to assess the relevance of such traits outside of the key's context, however.

A particularly potent application of the Clavis format will be an implementation in tandem with automated image recognition. Since statements are stored as separate entities, there is no fixed path through the key requiring that taxa are evaluated against their characteristics in any particular order. This means that the key can be applied to any subset of the taxa as easily as to the full taxon set. Which characters are displayed to the user will automatically adjust according to the subset of taxa. This allows for a reduction of the probable identifications by a machine learning algorithm as a first step, followed by keying of the relevant subset of taxa to make a final determination. This mechanism potentially reduces the need for much of the user input of a full key, thus saving the user considerable effort and reducing the possibility of the user providing incorrect input that is inherent each time user input is provided. Conversely, it reduces the reliability solely on machine learning for identification, providing a mechanism of quality control of the algorithm output, and an opportunity for the end user to learn a great deal more than they would with only a recognition model prediction.

The development of Clavis has been done in close collaboration with taxonomic experts. While this has enabled us to include many diverse features covering needs that have arisen in the past, no such endeavor can expect to produce a final version covering all needs. Further adoption may also bring to light ambiguities or shortcomings that will need to be addressed. Our aim is to continue to update the format, releasing new iterations with improvements. Use of previous versions will remain possible, and we aim to maintain backwards compatibility wherever possible. We invite the community to contribute to the development of Clavis, through submitting issues on GitHub, and resolving issues by answering questions or proposing code changes through pull requests. We hope that solutions supporting Clavis, be it key building software or end-user interfaces, both of which we plan to create examples of, will be shared openly as part of a broader ecosystem of use and re-use.

Our aim is for Clavis to be a relevant tool for storing the taxonomic knowledge needed for identification in a way that allows for the representation of the complexity and nuances inherent to such knowledge. The open exchange of taxonomic knowledge, unambiguously captured with as much of the auxiliary details needed for its application, is essential for the preservation of invaluable, increasingly elusive knowledge.

We believe that the storage of taxonomic knowledge with the level of detail exemplified here, combined with user interfaces making it accessible, is vital in enabling observers to gather the data needed for apt nature management. Particularly within citizen science, the potential of tools built on an identification key format such as Clavis is considerable. The more accessible this expert knowledge is, the more accurate the identifications made by the user will be, and keys on generally less well-known taxa can aid in closing the taxonomic gaps in the

data corpus. The possibility of storing the user input together with a reported observation, can provide important metadata on the identification and its quality. The results of these projected advances in data collection quality feed back into the areas where these data are used, from research and spatial distribution models to the decision making processes related to the biodiversity crisis in a changing world.

Data availability and licensing

The latest formal definition of Clavis can be found at `https://github.com/Artsdatabanken/Clavis`. All files related to this manuscript can be found at `https://doi.org/10.5281/zenodo.6585092`.

Figures 1-4 are all made by Wouter Koch and licensed under a CC BY 4.0 license. In Fig. 5, character illustrations are made by Hallvard Elven and licensed under a CC BY 4.0 license, while taxon images are made by Göran Liljeberg and licensed under a CC BY-SA 4.0 license.

Acknowledgements

We are grateful to Askild Aaberg Hofsøy Olsen for his feedback on the technical aspects of the Clavis schema.

References

1. Ceballos, G. *et al.* Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances* **1.** `https://doi.org/10.1126/sciadv.1400253` (June 2015).
2. Johnson, C. N. *et al.* Biodiversity losses and conservation responses in the Anthropocene. *Science* **356,** 270–275. `https://doi.org/10.1126/science.aam9317` (Apr. 2017).
3. Engel, M. S. *et al.* The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society* **193,** 381–387. `https://doi.org/10.1093/zoolinnean/zlab072` (Sept. 2021).
4. Silvertown, J. A new dawn for citizen science. *Trends in Ecology & Evolution* **24,** 467–471. `https://doi.org/10.1016/j.tree.2009.03.017` (Sept. 2009).
5. GBIF.org. *GBIF homepage* `https://www.gbif.org`.
6. GBIF.org. *Taxonomic distribution of occurrences* `https://www.gbif.org/occurrence/taxonomy?occurrence_status=present` (2022).

7. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7.** https://doi.org/10.1038/s41598-017-09084-6 (Aug. 2017).
8. Dobson, A. *et al.* Making Messy Data Work for Conservation. *One Earth* **2,** 455–465. ISSN: 2590-3322. https://doi.org/10.1016/j.oneear.2020.04.012 (2020).
9. Ebach, M. C., Valdecasas, A. G. & Wheeler, Q. D. Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics* **27,** 550–557. https://doi.org/10.1111/j.1096-0031.2011.00348.x (Feb. 2011).
10. Crall, A. W. *et al.* Assessing citizen science data quality: an invasive species case study. *Conservation Letters* **4,** 433–442. https://doi.org/10.1111/j.1755-263x.2011.00196.x (Aug. 2011).
11. Burgess, H. *et al.* The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation* **208,** 113–120. https://doi.org/10.1016/j.biocon.2016.05.014 (Apr. 2017).
12. Callaghan, C. T. *et al.* Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience* **71,** 55–63. ISSN: 0006-3568. https://doi.org/10.1093/biosci/biaa131 (Nov. 2020).
13. Bayraktarov, E. *et al.* Do Big Unstructured Biodiversity Data Mean More Knowledge? *Frontiers in Ecology and Evolution* **6.** https://doi.org/10.3389/fevo.2018.00239 (Jan. 2019).
14. Boakes, E. H. *et al.* Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific Reports* **6.** https://doi.org/10.1038/srep33051 (Sept. 2016).
15. Dallwitz, M. J. A general system for coding taxonomic descriptions. *TAXON* **29,** 41–46. https://doi.org/10.2307/1219595 (Feb. 1980).
16. Identic Pty Ltd. *Lucidcentral.org* https://www.lucidcentral.org (2022).
17. Fandom Community. *Pokémon GO Wiki* https://pokemongo.fandom.com/wiki/List_of_Pok%C3%A9mon (2022).
18. Crockford, D. & Morningstar, C. *Standard ECMA-404 The JSON Data Interchange Syntax* en. Tech. rep. (2017). https://doi.org/10.13140/RG.2.2.28181.14560.
19. Wright, A., Andrews, H., Hutton, B. & Dennis, G. *JSON Schema: A Media Type for Describing JSON Documents* https://datatracker.ietf.org/doc/html/draft-bhutton-json-schema-00 (2022).

Supporting information

S1 Appendix. Clavis JSON-schema. The formal definition of what constitutes a Clavis-compliant key.

S2 Appendix. Pokémon key example. A Clavis-compliant key to a number of Pokémon. Serves to illustrate all the different aspects that Clavis supports, rather than to provide a fully functional and complete key.

S3 Appendix. Titmice key example. A Clavis-compliant key to Norway's titmice (Paridae). Serves as a real life example of a fully functional and complete key, using only a selection of Clavis' capabilities.

Clavis: an open and versatile identification key format

SUPPLEMENTARY MATERIALS

S1 Appendix. Clavis JSON-schema

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Clavis identification key schema",
  "description": "Clavis-compliant keys contain knowledge that may be
used to distinguish taxa from each other.",
  "type": "object",
  "required": [
    "$schema",
    "title",
    "language",
    "license",
    "creator",
    "lastModified",
    "identifier",
    "taxa",
    "characters",
    "statements",
    "persons"
  ],
  "properties": {
    "$schema": {
      "description": "The schema url of (this) schema defining the
format of the key.",
      "$ref": "#/definitions/url"
    },
    "title": {
      "description": "The name of the key",
      "comment": "Accepts array for multilingual support.",
      "$ref": "#/definitions/localizedString",
      "examples": [
        "Birds of Norway"
      ]
    },
```

```
    "media": {
      "description": "The logo/illustration image of the key.",
      "$ref": "#/definitions/mediaID"
    },
    "description": {
      "description": "Short description of the key (valid
markdown).",
      "comment": "Accepts array for multilingual support.",
      "$ref": "#/definitions/localizedString",
      "contentMediaType": "text/markdown"
    },
    "descriptionDetails": {
      "description": "Extended description of the key that supplements
the description (valid markdown).",
      "comment": "Accepts array for multilingual support.",
      "$ref": "#/definitions/localizedString",
      "contentMediaType": "text/markdown"
    },
    "descriptionUrl": {
      "description": "Hyperlink to more information on the key (valid
url).",
      "comment": "Accepts array for multilingual support.",
      "$ref": "#/definitions/localizedUrl"
    },
    "audience": {
      "description": "Description of the intended audience for the
key.",
      "comment": "Accepts array for multilingual support.",
      "$ref": "#/definitions/localizedString",
      "examples": [
        "Undergraduate students and up."
      ]
    },
    "source": {
      "description": "Source of the key.",
      "comment": "Accepts array for multilingual support.",
      "$ref": "#/definitions/localizedString",
      "examples": [
        "Koch, Wouter (2019). Birds of Norway. ISBN 1234567890"
      ]
    },
    "sourceUrl": {
      "description": "Hyperlink to the source of the key (valid
url).",
      "comment": "Accepts array for multilingual support.",
      "$ref": "#/definitions/localizedUrl",
      "examples": [
        "https://doi.org/10.1126/science.1251554"
```

```
      ]
    },
    "geography": {
      "description": "The region for which the key is valid (e.g.
covers all subtaxa), represented as a geography object.",
      "$ref": "#/definitions/geography"
    },
    "primaryContact": {
      "description": "The organization- or person-id that is the main
contact point for the key.",
      "oneOf": [
        {
          "$ref": "#/definitions/personID"
        },
        {
          "$ref": "#/definitions/organizationID"
        }
      ]
    },
    "creator": {
      "description": "The id(s) of the creator(s) of the key",
      "oneOf": [
        {
          "$ref": "#/definitions/personID"
        },
        {
          "type": "array",
          "items": {
            "$ref": "#/definitions/personID"
          }
        }
      ]
    },
    "contributor": {
      "description": "The id(s) of the contributor(s) of the key",
      "oneOf": [
        {
          "$ref": "#/definitions/personID"
        },
        {
          "type": "array",
          "items": {
            "$ref": "#/definitions/personID"
          }
        }
      ]
    },
    "publisher": {
```

```
      "description": "The id(s) of the publishing institutions of the
key.",
      "oneOf": [
        {
          "$ref": "#/definitions/organizationID"
        },
        {
          "type": "array",
          "items": {
            "$ref": "#/definitions/organizationID"
          }
        }
      ]
    },
    "license": {
      "description": "The url to the license under which the key
falls.",
      "$ref": "#/definitions/url",
      "examples": [
        "https://creativecommons.org/licenses/by/4.0/"
      ]
    },
    "language": {
      "description": "The ISO 639-1 code(s) of the key language(s).",
      "comment": "String for a single language, array of strings for
multilingual support. If used as an array, be sure to use the
localizedString and localizedUrl as arrays too.",
      "oneOf": [
        {
          "type": "string",
          "pattern": "^[a-z]{2}$"
        },
        {
          "type": "array",
          "items": {
            "type": "string",
            "pattern": "^[a-z]{2}$"
          }
        }
      ],
      "examples": [
        "en",
        "nb",
        [
          "en",
          "nb"
        ]
      ]
```

```
    },
    "created": {
      "description": "The moment the key was made or first published,
as 'YYYY-MM-DD hh:mm:ss'.",
      "type": "string",
      "pattern": "^20\d\d-(0[1-9]|1[0-2])-([012]\d|3[01])
([01]\d|2[0-3]):([0-5]\d):([0-5]\d)$",
      "examples": [
        "2019-05-21 22:51:55"
      ]
    },
    "lastModified": {
      "description": "The most recent moment the key was modified, as
'YYYY-MM-DD hh:mm:ss'.",
      "type": "string",
      "pattern": "^20\d\d-(0[1-9]|1[0-2])-([012]\d|3[01])
([01]\d|2[0-3]):([0-5]\d):([0-5]\d)$",
      "examples": [
        "2019-05-21 22:51:55"
      ]
    },
    "identifier": {
      "description": "The GUID of this key (persistent regardless of
version).",
      "type": "string"
    },
    "url": {
      "description": "The url of where the key lives (to check for
newer versions).",
      "$ref": "#/definitions/url"
    },
    "externalServices": {
      "description": "Services used by the key for lookups of images,
taxa, etc.",
      "type": "array",
      "items": {
        "$ref": "#/definitions/externalService"
      }
    },
    "userRequirements": {
      "description": "Requirements to the users of the various
characters, so that the user can be warned, helped, etc.",
      "type": "array",
      "items": {
        "$ref": "#/definitions/userRequirement"
      }
    },
    "taxa": {
```

```
      "description": "Taxa (e.g. species) the key can resolve to. Do
not have to be exclusively taxonomic units.",
      "comment": "Taxa to which the key can resolve (either the taxa
directly or their children).",
      "type": "array",
      "items": {
        "$ref": "#/definitions/taxon"
      }
    },
    "characters": {
      "description": "Characters (questions, e.g. 'Wing color' or
'Number of spots') used to distinguish between two or more taxa.",
      "type": "array",
      "items": {
        "$ref": "#/definitions/character"
      }
    },
    "statements": {
      "description": "Relationships between taxa and character states
(or lack thereof) that define those taxa.",
      "type": "array",
      "items": {
        "$ref": "#/definitions/statement"
      }
    },
    "persons": {
      "description": "Persons that are connected to (parts of) the
key, such as creators.",
      "type": "array",
      "items": {
        "$ref": "#/definitions/person"
      }
    },
    "organizations": {
      "description": "Organizations that are connected to (parts of)
the key or persons, such as employers and publishers.",
      "type": "array",
      "items": {
        "$ref": "#/definitions/organization"
      }
    },
    "mediaElements": {
      "description": "Media elements that are used in the key.",
      "type": "array",
      "items": {
        "$ref": "#/definitions/localizedMediaElement"
      }
    }
```

```
  },
  "additionalProperties": false,
  "definitions": {
    "localizedString": {
      "description": "Language-dependent string or object of strings,
with keys corresponding to the languages supported by the key.",
      "oneOf": [
        {
          "type": "string"
        },
        {
          "type": "object",
          "propertyNames": {
            "pattern": "^[a-z]{2}$"
          },
          "properties": {},
          "additionalProperties": {
            "type": "string"
          }
        }
      ]
    },
    "localizedUrl": {
      "description": "Language-dependent urls or object of urls,
corresponding to the languages supported by the key.",
      "oneOf": [
        {
          "$ref": "#/definitions/url"
        },
        {
          "type": "object",
          "propertyNames": {
            "pattern": "^[a-z]{2}$"
          },
          "properties": {},
          "additionalProperties": {
            "$ref": "#/definitions/url"
          }
        }
      ]
    },
    "localizedMediaElement": {
      "type": "object",
      "description": "Language-dependent media element or object of
media elements, corresponding to the languages supported by the
key.",
      "properties": {
        "id": {
```

```
322           "description": "Internally unique id of the localized media
323 element.",
324           "$ref": "#/definitions/mediaID"
325         },
326         "mediaElement": {
327           "description": "The media element or media elements (one for
328 each language).",
329           "oneOf": [
330             {
331               "$ref": "#/definitions/mediaElement"
332             },
333             {
334               "type": "object",
335               "propertyNames": {
336                 "pattern": "^[a-z]{2}$"
337               },
338               "properties": {},
339               "additionalProperties": {
340                 "$ref": "#/definitions/mediaElement"
341               }
342             }
343           ]
344         }
345       },
346       "additionalProperties": false
347     },
348     "url": {
349       "description": "String formed as a url, or an external
350 resource.",
351       "oneOf": [
352         {
353           "type": "string",
354           "format": "uri"
355         },
356         {
357           "$ref": "#/definitions/externalResource"
358         }
359       ]
360     },
361     "taxonID": {
362       "description": "String used as an internal ID for a taxon.
363 Lowercase alphanumeric and underscores are allowed.",
364       "type": "string",
365       "pattern": "^taxon:[a-z0-9_]+$"
366     },
367     "characterID": {
368       "description": "String used as an internal ID for a character.
369 Lowercase alphanumeric and underscores are allowed.",
```

```
      "type": "string",
      "pattern": "^character:[a-z0-9_]+$"
    },
    "stateID": {
      "description": "String used as an internal ID for a state.
Lowercase alphanumeric and underscores are allowed.",
      "type": "string",
      "pattern": "^state:[a-z0-9_]+$"
    },
    "personID": {
      "description": "String used as an internal ID for a person.
Lowercase alphanumeric and underscores are allowed.",
      "type": "string",
      "pattern": "^person:[a-z0-9_]+$"
    },
    "organizationID": {
      "description": "String used as an internal ID for an
organization. Lowercase alphanumeric and underscores are allowed.",
      "type": "string",
      "pattern": "^organization:[a-z0-9_]+$"
    },
    "serviceID": {
      "description": "String used as an internal ID for a service.
Lowercase alphanumeric and underscores are allowed.",
      "type": "string",
      "pattern": "^service:[a-z0-9_]+$"
    },
    "statementID": {
      "description": "String used as an internal ID for a statement.
Lowercase alphanumeric and underscores are allowed.",
      "type": "string",
      "pattern": "^statement:[a-z0-9_]+$"
    },
    "userRequirementID": {
      "description": "String used as an internal ID for a user
requirement. Lowercase alphanumeric and underscores are allowed.",
      "type": "string",
      "pattern": "^requirement:[a-z0-9_]+$"
    },
    "mediaID": {
      "description": "String used as an internal ID for a media
element. Lowercase alphanumeric and underscores are allowed.",
      "type": "string",
      "pattern": "^media:[a-z0-9_]+$"
    },
    "mediaFile": {
      "type": "object",
      "properties": {
```

```
        "title": {
          "description": "The title of the media file.",
          "$ref": "#/definitions/localizedString"
        },
        "url": {
          "description": "The reference to the media file (url or
resource).",
          "$ref": "#/definitions/url"
        },
        "file": {
          "description": "The actual media file (base64 or svg) as a
data URI scheme.",
          "oneOf": [
            {
              "type": "string",
              "pattern":
"^data:([a-z0-9/]+)base64,([a-zA-Z0-9+/=]+)$"
            },
            {
              "type": "string",
              "pattern": "^data:image/svg\+xml;utf8,(.*)$"
            }
          ]
        },
        "width": {
          "description": "The number of pixels horizontally (if a
bitmap image or video).",
          "type": "integer"
        },
        "height": {
          "description": "The number of pixels vertically (if a bitmap
image or video).",
          "type": "integer"
        },
        "length": {
          "description": "The length in seconds of an audio or video
file.",
          "type": "integer"
        },
        "placeholder": {
          "description": "Image file that can be shown instead of the
video or audio file.",
          "$ref": "#/definitions/mediaID"
        },
        "creator": {
          "description": "The id(s) of the creator(s) of the media
file",
          "oneOf": [
```

```
            {
              "$ref": "#/definitions/personID"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/personID"
              }
            }
          ]
        },
        "contributor": {
          "description": "The id(s) of the contributor(s) of the media
file",
          "oneOf": [
            {
              "$ref": "#/definitions/personID"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/personID"
              }
            }
          ]
        },
        "publisher": {
          "description": "The id(s) of the publishing institutions of
the media file.",
          "oneOf": [
            {
              "$ref": "#/definitions/organizationID"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/organizationID"
              }
            }
          ]
        },
        "license": {
          "description": "The url to the license under which the media
file falls.",
          "$ref": "#/definitions/url",
          "examples": [
            "https://creativecommons.org/licenses/by/4.0/"
          ]
```

```
514         }
515       },
516       "additionalProperties": false
517     },
518     "mediaElement": {
519       "description": "A media element (collection of various formats
520 of the same media object).",
521       "type": "object",
522       "properties": {
523         "file": {
524           "description": "The various formats of the same media
525 object.",
526           "oneOf": [
527             {
528               "$ref": "#/definitions/mediaFile"
529             },
530             {
531               "type": "array",
532               "items": {
533                 "$ref": "#/definitions/mediaFile"
534               }
535             }
536           ]
537         }
538       },
539       "additionalProperties": false
540     },
541     "multiPolygon": {
542       "description": "The coordinates array of a GeoJSON
543 MultiPolygon.",
544       "type": "array",
545       "items": {
546         "type": "array",
547         "items": {
548           "type": "array",
549           "items": {
550             "type": "array",
551             "items": {
552               "type": "number"
553             }
554           }
555         }
556       }
557     },
558     "geography": {
559       "description": "A geographic element (name, polygon, and/or
560 external service).",
561       "type": "object",
```

```
      "properties": {
        "name": {
          "description": "The name of the area(s).",
          "comment": "Accepts array for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "examples": [
            "Norway",
            "Europe",
            "Trøndelag",
            [
              "Norge",
              "Norway"
            ]
          ]
        },
        "polygon": {
          "description": "The geographical area(s), represented as the
coordinates array of a GeoJSON MultiPolygon.",
          "$ref": "#/definitions/multiPolygon"
        },
        "service": {
          "description": "An url or external service that returns
geographical information.",
          "$ref": "#/definitions/url"
        }
      },
      "additionalProperties": false
    },
    "externalResource": {
      "description": "A resource managed elsewhere.",
      "type": "object",
      "properties": {
        "serviceId": {
          "description": "The id to one of the externalServices
defined.",
          "$ref": "#/definitions/serviceID"
        },
        "externalId": {
          "description": "The id of the resource at the
externalService.",
          "type": "string"
        }
      },
      "additionalProperties": false
    },
    "externalService": {
      "description": "Service used by the key, for media files,
taxonomy and/or nomenclature, species distributions, etc.",
```

```
610       "type": "object",
611       "required": [
612         "id"
613       ],
614       "properties": {
615         "id": {
616           "description": "Internally unique id to the service.",
617           "$ref": "#/definitions/serviceID"
618         },
619         "title": {
620           "description": "Name of the service.",
621           "type": "string"
622         },
623         "description": {
624           "description": "Description of the service.",
625           "type": "string"
626         },
627         "provider": {
628           "description": "Provider of the service.",
629           "type": "string"
630         },
631         "url": {
632           "description": "Url for the service documentation.",
633           "$ref": "#/definitions/url"
634         }
635       },
636       "additionalProperties": false
637     },
638     "person": {
639       "type": "object",
640       "required": [
641         "id",
642         "name"
643       ],
644       "properties": {
645         "id": {
646           "$ref": "#/definitions/personID"
647         },
648         "name": {
649           "description": "Full name of the person",
650           "comment": "Accepts object for multilingual support.",
651           "$ref": "#/definitions/localizedString"
652         },
653         "email": {
654           "description": "Email address of the person",
655           "type": "string",
656           "format": "email"
657         },
```

```
        "url": {
          "description": "Hyperlink to more information on the person
(valid url).",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedUrl"
        },
        "media": {
          "description": "A media file (image) representing the
person.",
          "$ref": "#/definitions/mediaID"
        },
        "affiliation": {
          "description": "Organization id(s) the person is affiliated
with.",
          "oneOf": [
            {
              "$ref": "#/definitions/organizationID"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/organizationID"
              }
            }
          ]
        }
      },
      "additionalProperties": false
    },
    "organization": {
      "type": "object",
      "required": [
        "id",
        "name"
      ],
      "properties": {
        "id": {
          "$ref": "#/definitions/organizationID"
        },
        "name": {
          "description": "Name of the organization",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString"
        },
        "url": {
          "description": "Hyperlink to more information on the
organization (valid url).",
          "comment": "Accepts object for multilingual support.",
```

```
706           "$ref": "#/definitions/localizedUrl"
707         },
708         "primaryContact": {
709           "description": "The person-id that is the main contact point
710 for the organization.",
711           "$ref": "#/definitions/personID"
712         },
713         "media": {
714           "description": "A media file (image) representing the
715 organization, such as a logo.",
716           "$ref": "#/definitions/mediaID"
717         }
718       },
719       "additionalProperties": false
720     },
721     "userRequirement": {
722       "type": "object",
723       "required": [
724         "id"
725       ],
726       "properties": {
727         "id": {
728           "$ref": "#/definitions/userRequirementID"
729         },
730         "title": {
731           "comment": "Accepts array for multilingual support.",
732           "$ref": "#/definitions/localizedString"
733         },
734         "warning": {
735           "comment": "Accepts array for multilingual support.",
736           "$ref": "#/definitions/localizedString"
737         },
738         "description": {
739           "description": "Short description of the requirements to the
740 user (valid markdown).",
741           "comment": "Accepts object for multilingual support.",
742           "$ref": "#/definitions/localizedString",
743           "contentMediaType": "text/markdown"
744         },
745         "descriptionDetails": {
746           "description": "Extended description of the requirements to
747 the user that supplements the description (valid markdown).",
748           "comment": "Accepts object for multilingual support.",
749           "$ref": "#/definitions/localizedString",
750           "contentMediaType": "text/markdown"
751         },
752         "descriptionUrl": {
753           "description": "Hyperlink to more information on the
```

```
754 requirements to the user (valid url).",
755           "comment": "Accepts object for multilingual support.",
756           "$ref": "#/definitions/localizedUrl"
757         },
758         "media": {
759           "description": "Media or illustration that informs the user
760 on the requirements to the user.",
761           "comment": "Accepts object for multilingual support.",
762           "$ref": "#/definitions/mediaID"
763         }
764       },
765       "additionalProperties": false
766     },
767     "taxon": {
768       "type": "object",
769       "oneOf": [
770         {
771           "required": [
772             "id",
773             "scientificName"
774           ]
775         },
776         {
777           "required": [
778             "id",
779             "externalReference"
780           ]
781         },
782         {
783           "required": [
784             "id",
785             "label"
786           ]
787         }
788       ],
789       "properties": {
790         "id": {
791           "description": "Internally unique id to the taxon.",
792           "$ref": "#/definitions/taxonID"
793         },
794         "scientificName": {
795           "description": "Scientific name of the taxon.",
796           "minLength": 5,
797           "type": "string",
798           "examples": [
799             "Vulpes lagopus"
800           ]
801         },
```

```
        "scientificNameAuthor": {
          "description": "Author string of the scientific name of the
taxon.",
          "type": "string",
          "examples": [
            "Koch, 1888"
          ]
        },
        "placeholderName": {
          "description": "Name that can be shown while fetching the
name externally. Also useful for editing the key.",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "examples": [
            "B. hortorum (melanistic queen)"
          ]
        },
        "vernacularName": {
          "description": "Vernacular name of the taxon.",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "examples": [
            "fjellrev",
            {
              "no": "fjellrev",
              "en": "Arctic Fox"
            }
          ]
        },
        "media": {
          "description": "Media elements of the taxon.",
          "oneOf": [
            {
              "$ref": "#/definitions/mediaID"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/mediaID"
              }
            }
          ]
        },
        "description": {
          "description": "Short description of the taxon (valid
markdown).",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
```

```
          "contentMediaType": "text/markdown"
        },
        "descriptionDetails": {
          "description": "Extended description of the taxon that
supplements the description (valid markdown).",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "contentMediaType": "text/markdown"
        },
        "descriptionUrl": {
          "description": "Hyperlink or resource to more information on
the taxon.",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedUrl"
        },
        "rank": {
          "description": "Name of the level of the taxon.",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "examples": [
            "slekt",
            {
              "no": "slekt",
              "en": "genus"
            }
          ]
        },
        "label": {
          "description": "Type of morph of the taxon.",
          "type": "string",
          "minLength": 0,
          "examples": [
            "male",
            "♀",
            "larva"
          ]
        },
        "isEndPoint": {
          "description": "Whether the key should stop when this taxon
is the only remaining possibility, even when it has multiple children
remaining.",
          "comment": "Default is FALSE (if not specified). A taxon
without children is alway an endpoint by definition, unless one of its
ancestors overrides this by being specified as an endpoint.",
          "type": "boolean"
        },
        "children": {
          "type": "array",
```

```
          "items": {
            "$ref": "#/definitions/taxon"
          }
        },
        "externalReference": {
          "description": "Reference to a taxon at one or more
providers, each as an object with a provider id and a taxon id at that
provider.",
          "comment": "Accepts an array for multiple sources. Each
element accepts an object for multilingual support.",
          "oneOf": [
            {
              "$ref": "#/definitions/localizedUrl"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/localizedUrl"
              }
            }
          ]
        },
        "followUp": {
          "description": "Url or reference to instance at external
service for a key for this taxon, that for instance can be used to
identify to a lower rank than the current key can.",
          "comment": "Accepts array for multilingual support.",
          "$ref": "#/definitions/localizedUrl"
        },
        "geography": {
          "description": "The area(s) in which the taxon occurs,
represented as a geography object.",
          "$ref": "#/definitions/geography"
        }
      },
      "additionalProperties": false
    },
    "character": {
      "type": "object",
      "oneOf": [
        {
          "required": [
            "id",
            "title",
            "states"
          ]
        },
        {
```

```
946           "required": [
947             "id",
948             "title",
949             "type",
950             "min",
951             "max",
952             "stepSize",
953             "unit"
954           ]
955         }
956       ],
957       "properties": {
958         "id": {
959           "description": "Internally unique id to the character.",
960           "$ref": "#/definitions/characterID"
961         },
962         "title": {
963           "description": "Name of the character.",
964           "comment": "Accepts array for multilingual support.",
965           "$ref": "#/definitions/localizedString",
966           "examples": [
967             "Color of the wings"
968           ]
969         },
970         "media": {
971           "description": "The media element(s) of the character. Can
972 be used to inform user of relevant structures etc.",
973           "oneOf": [
974             {
975               "$ref": "#/definitions/mediaID"
976             },
977             {
978               "type": "array",
979               "items": {
980                 "$ref": "#/definitions/mediaID"
981               }
982             }
983           ]
984         },
985         "description": {
986           "description": "Short description of the character (valid
987 markdown).",
988           "comment": "Accepts object for multilingual support.",
989           "$ref": "#/definitions/localizedString",
990           "contentMediaType": "text/markdown"
991         },
992         "descriptionDetails": {
993           "description": "Extended description of the character that
```

```
supplements the description (valid markdown).",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "contentMediaType": "text/markdown"
        },
        "descriptionUrl": {
          "description": "Hyperlink or resource to more information on
the character.",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedUrl"
        },
        "type": {
          "description": "Type of the character (exclusive when states
are categorical and mutually exclusive, non-exclusive when these are
non-exclusive, or numerical when the state is numerical).",
          "comment": "Default is exclusive (if not specified).",
          "type": "string",
          "enum": [
            "exclusive",
            "non-exclusive",
            "numerical"
          ]
        },
        "userRequirement": {
          "description": "Id to the userRequirement required to answer
this character.",
          "comment": "Has to be one of the userRequirement defined on
the key level.",
          "$ref": "#/definitions/userRequirementID"
        },
        "logicalPremise": {
          "description": "Logical requirement that has to be fulfilled
for this question to be asked.",
          "comment": "Has to refer to stateIds, that have to be fully
true (either answered or all alternatives ruled out). Can use !, &&,
||, (, ), <, >, =.",
          "type": "string",
          "pattern": "^(( && )|( \|\|
)|(&&)|(\|\|)|[a-z0-9_:()!<>=])+$"
        },
        "min": {
          "type": "number",
          "description": "The minimum numerical value for the
character."
        },
        "max": {
          "type": "number",
          "description": "The maximum numerical value for the
```

```
character."
        },
        "stepSize": {
          "type": "number",
          "description": "The increments with which the numerical
value of the character can be specified."
        },
        "unit": {
          "description": "The unit of the numerical value.",
          "$ref": "#/definitions/localizedString",
          "examples": [
            "mm",
            "meters below the surface",
            "spots",
            "legs",
            "kg"
          ]
        },
        "states": {
          "oneOf": [
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/state"
              }
            },
            {
              "$ref": "#/definitions/state"
            }
          ]
        }
      },
      "additionalProperties": false
    },
    "state": {
      "description": "The value a character can have.",
      "type": "object",
      "required": [
        "id",
        "title"
      ],
      "properties": {
        "id": {
          "description": "Internally unique id of the state.",
          "$ref": "#/definitions/stateID"
        },
        "title": {
          "description": "Content of the state.",
```

```
          "comment": "Only to be used for categorical characters.
Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString"
        },
        "media": {
          "description": "Media element(s) that illustrate the
state.",
          "oneOf": [
            {
              "$ref": "#/definitions/mediaID"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/mediaID"
              }
            }
          ]
        },
        "description": {
          "description": "Short description of the state (valid
markdown).",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "contentMediaType": "text/markdown"
        },
        "descriptionDetails": {
          "description": "Extended description of the state that
supplements the description (valid markdown).",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "contentMediaType": "text/markdown"
        },
        "descriptionUrl": {
          "description": "Hyperlink or resource to more information on
the state.",
          "comment": "Accepts object for multilingual support.",
          "$ref": "#/definitions/localizedUrl"
        }
      },
      "additionalProperties": false
    },
    "statement": {
      "description": "A fact connecting a taxon and a character
through a certain value.",
      "type": "object",
      "required": [
        "id",
```

```
        "taxon",
        "character",
        "value",
        "frequency"
      ],
      "properties": {
        "id": {
          "description": "Internally unique id of the statement.",
          "$ref": "#/definitions/statementID"
        },
        "taxon": {
          "description": "Id of the taxon this statement is about.",
          "$ref": "#/definitions/taxonID"
        },
        "character": {
          "description": "Id of the character this statement is
about.",
          "$ref": "#/definitions/characterID"
        },
        "value": {
          "description": "A value for this character for this taxon.
Must be either the id of a state, or an array of floats [min, max] for
a numerical range.",
          "oneOf": [
            {
              "$ref": "#/definitions/stateID"
            },
            {
              "type": "array",
              "items": {
                "type": "number"
              },
              "minItems": 2,
              "maxItems": 2
            }
          ]
        },
        "frequency": {
          "description": "The frequency with which the taxon has this
value for this character.",
          "type": "number",
          "minimum": 0,
          "maximum": 1
        },
        "geography": {
          "description": "The area(s) in which the taxon can have this
property, represented as a geography object.",
          "$ref": "#/definitions/geography"
```

```
        },
        "media": {
          "description": "Illustration(s) of this particular taxon
having this particular property (this value for this character).",
          "oneOf": [
            {
              "$ref": "#/definitions/mediaID"
            },
            {
              "type": "array",
              "items": {
                "$ref": "#/definitions/mediaID"
              }
            }
          ]
        },
        "description": {
          "description": "Short description of the taxon having this
property (valid markdown).",
          "comment": "Accepts array for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "contentMediaType": "text/markdown"
        },
        "descriptionDetails": {
          "description": "Extended description of the taxon having
this property that supplements the description (valid markdown).",
          "comment": "Accepts array for multilingual support.",
          "$ref": "#/definitions/localizedString",
          "contentMediaType": "text/markdown"
        },
        "descriptionUrl": {
          "description": "Hyperlink or resource to more information on
the taxon having this property.",
          "comment": "Accepts array for multilingual support.",
          "$ref": "#/definitions/localizedUrl"
        }
      },
      "additionalProperties": false
    }
  }
}
```

S2 Appendix. Pokémon key example

```
{
  "$schema":
"https://raw.githubusercontent.com/WouterKoch/Clavis/main/Schema/Clavi
s.json",
  "title": "A key to a selection of Pokémon",
  "language": "en",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "creator": "person:wouterkoch",
  "lastModified": "2022-03-19 21:35:25",
  "identifier": "26b57071-15ca-4b44-92a4-b61181f15373",
  "persons": [
    {
      "id": "person:wouterkoch",
      "name": "Wouter Koch"
    }
  ],
  "organizations": [
    {
      "id": "organization:ntnu",
      "name": "Norwegian University of Science and Technology",
      "url": "https://www.ntnu.no"
    }
  ],
  "taxa": [
    {
      "id": "taxon:pikachuidae",
      "scientificName": "Pikachuidae",
      "children": [
        {
          "id": "taxon:pokemon_172",
          "scientificName": "Pichu",
          "children": [
            {
              "id": "taxon:pokemon_172_standard",
              "label": ""
            },
            {
              "id": "taxon:pokemon_172_shiny",
              "label": "Shiny"
            }
          ]
        },
        {
          "id": "taxon:pokemon_025",
          "scientificName": "Pikachu",
          "externalReference": [
```

```
            {
              "serviceId": "service:wikidata",
              "externalId": "Q9351"
            },
            {
              "serviceId": "service:example_api",
              "externalId": "Pikachu"
            }
          ],
          "followUp": "https://example.com/pikachu_costumes",
          "children": [
            {
              "id": "taxon:pokemon_025_standard",
              "isEndPoint": true,
              "label": "",
              "children": [
                {
                  "id": "taxon:pokemon_025_standard_male",
                  "label": "♂"
                },
                {
                  "id": "taxon:pokemon_025_standard_female",
                  "label": "♀"
                }
              ]
            },
            {
              "id": "taxon:pokemon_025_shiny",
              "label": "Shiny"
            }
          ]
        },
        {
          "id": "taxon:pokemon_026",
          "scientificName": "Raichu",
          "children": [
            {
              "id": "taxon:pokemon_026_standard",
              "label": ""
            },
            {
              "id": "taxon:pokemon_026_shiny",
              "label": "Shiny"
            }
          ]
        }
      ]
    },
```

```
    {
      "id": "taxon:pokemon_115",
      "scientificName": "Kangaskhan",
      "geography": {
        "polygon": [
          [
            [
              [
                130.0341796875,
                -10.228437266155943
              ],
              [
                111.97265625,
                -21.779905342529634
              ],
              [
                115.09277343749999,
                -36.91476428895593
              ],
              [
                131.2646484375,
                -32.916485347314385
              ],
              [
                141.8994140625,
                -40.44694705960048
              ],
              [
                150.82031249999997,
                -38.8225909761771
              ],
              [
                154.95117187499997,
                -26.15543796871355
              ],
              [
                142.3388671875,
                -10.09867012060338
              ],
              [
                138.69140625,
                -12.382928338487396
              ],
              [
                130.0341796875,
                -10.228437266155943
              ]
            ]
```

```
          ]
        ]
      }
    },
    {
      "id": "taxon:castform",
      "scientificName": "Castform"
    }
  ],
  "characters": [
    {
      "id": "character:type_of_pokemon",
      "title": "Pokémon type",
      "states": [
        {
          "id": "state:pokemon_type_electric",
          "title": "Electric"
        }
      ]
    },
    {
      "id": "character:color",
      "title": "Color of body",
      "states": [
        {
          "id": "state:color_blue",
          "title": "Blue"
        },
        {
          "id": "state:color_red",
          "title": "Red"
        }
      ]
    },
    {
      "id": "character:all_colors",
      "title": "Colors on body of the Pokémon",
      "type": "non-exclusive",
      "states": [
        {
          "id": "state:colors_yellow",
          "title": "Yellow"
        },
        {
          "id": "state:colors_red",
          "title": "Red"
        },
        {
```

```
          "id": "state:colors_black",
          "title": "Black"
        },
        {
          "id": "state:colors_brown",
          "title": "Brown"
        },
        {
          "id": "state:colors_white",
          "title": "White"
        }
      ]
    },
    {
      "id": "character:tail_shape",
      "title": "Shape of the tail end",
      "states": [
        {
          "id": "state:pointy_tail",
          "title": "Pointy",
          "media": "media:pointy_tail"
        },
        {
          "id": "state:lobed_tail",
          "title": "Double-lobed",
          "media": "media:lobed_tail"
        }
      ]
    },
    {
      "id": "character:wearing_hat",
      "title": "Is the Pokémon wearing a hat?",
      "states": [
        {
          "id": "state:hat",
          "title": "Yes"
        },
        {
          "id": "state:no_hat",
          "title": "No"
        }
      ]
    },
    {
      "id": "character:hat_shape",
      "title": "What is the style of the hat?",
      "logicalPremise": "state:hat",
      "states": [
```

```
        {
          "id": "state:bowler_hat",
          "title": "Bowler hat"
        },
        {
          "id": "state:top_hat",
          "title": "Top hat"
        }
      ]
    },
    {
      "id": "character:weight",
      "title": "How much does the Pokémon weigh?",
      "userRequirement": "requirement:catch",
      "type": "numerical",
      "min": 1,
      "max": 125,
      "stepSize": 1,
      "unit": "kg"
    }
  ],
  "statements": [
    {
      "id": "statement:pikachu_is_electric",
      "taxon": "taxon:pokemon_025",
      "character": "character:type_of_pokemon",
      "value": "state:pokemon_type_electric",
      "frequency": 1
    },
    {
      "id": "statement:pikachu_is_never_blue",
      "taxon": "taxon:pokemon_025",
      "character": "character:color",
      "value": "state:color_blue",
      "frequency": 0
    },
    {
      "id": "statement:pikachu_lobed_tail",
      "taxon": "taxon:pokemon_025",
      "character": "character:tail_shape",
      "value": "state:lobed_tail",
      "frequency": 0.5
    },
    {
      "id": "statement:pikachu_pointy_tail",
      "taxon": "taxon:pokemon_025",
      "character": "character:tail_shape",
      "value": "state:pointy_tail",
```

```
      "frequency": 0.5
    },
    {
      "id": "statement:pikachu_weight",
      "taxon": "taxon:pokemon_025",
      "character": "character:weight",
      "value": [
        2.98,
        10.1
      ],
      "frequency": 1
    },
    {
      "id": "statement:castform_rain_type",
      "taxon": "taxon:castform",
      "character": "character:castform_type",
      "value": "state:rainy_castform",
      "frequency": 0.2
    },
    {
      "id": "statement:castform_rain_type_bergen",
      "taxon": "taxon:castform",
      "character": "character:castform_type",
      "value": "state:rainy_castform",
      "frequency": 0.9,
      "geography": {
        "polygon": [
          [
            [
              [
                5.27618408203125,
                60.44976847885747
              ],
              [
                5.218505859375,
                60.4233434866285
              ],
              [
                5.27618408203125,
                60.36160157353732
              ],
              [
                5.395660400390625,
                60.36839212633114
              ],
              [
                5.4052734375,
                60.421309904895715
```

```
              ],
              [
                5.27618408203125,
                60.44976847885747
              ]
            ]
          ]
        ]
      }
    },
    {
      "id": "statement:pikachu_contains_yellow",
      "taxon": "taxon:pokemon_025",
      "character": "character:all_colors",
      "value": "state:colors_yellow",
      "frequency": 1
    },
    {
      "id": "statement:pikachu_contains_black",
      "taxon": "taxon:pokemon_025",
      "character": "character:all_colors",
      "value": "state:colors_black",
      "frequency": 1
    },
    {
      "id": "statement:pikachu_contains_red",
      "taxon": "taxon:pokemon_025",
      "character": "character:all_colors",
      "value": "state:colors_red",
      "frequency": 1
    },
    {
      "id": "statement:pikachu_contains_no_brown",
      "taxon": "taxon:pokemon_025",
      "character": "character:all_colors",
      "value": "state:colors_brown",
      "frequency": 0
    },
    {
      "id": "statement:pikachu_contains_no_white",
      "taxon": "taxon:pokemon_025",
      "character": "character:all_colors",
      "value": "state:colors_white",
      "frequency": 0
    },
    {
      "id": "statement:raichu_contains_yellow",
      "taxon": "taxon:pokemon_026",
```

```
      "character": "character:all_colors",
      "value": "state:colors_yellow",
      "frequency": 1
    },
    {
      "id": "statement:raichu_contains_no_black",
      "taxon": "taxon:pokemon_026",
      "character": "character:all_colors",
      "value": "state:colors_black",
      "frequency": 0
    },
    {
      "id": "statement:raichu_contains_no_red",
      "taxon": "taxon:pokemon_026",
      "character": "character:all_colors",
      "value": "state:colors_red",
      "frequency": 0
    },
    {
      "id": "statement:raichu_contain_brown",
      "taxon": "taxon:pokemon_026",
      "character": "character:all_colors",
      "value": "state:colors_brown",
      "frequency": 1
    },
    {
      "id": "statement:raichu_contains_white",
      "taxon": "taxon:pokemon_026",
      "character": "character:all_colors",
      "value": "state:colors_white",
      "frequency": 1
    }
  ],
  "userRequirements": [
    {
      "id": "requirement:catch",
      "title": "Catching required",
      "warning": "To answer this, you have to catch the Pokémon
first.",
      "description": "1. Select a pokéball color.\n2. Hold the ball,
spinning it a few times.\n3. Fling the pokéball towards the
Pokémon, adjusting for the curveball generated from spinning the
ball.\n4. Try to hit the circle when it is at its smallest.",
      "descriptionUrl":
"https://niantic.helpshift.com/hc/en/6-pokemon-go/faq/102-finding-catc
hing-wild-pokemon/"
    }
  ],
```

```
  "externalServices": [
    {
      "id": "service:wikidata",
      "title": "Wikidata",
      "url": "https://www.wikidata.org/w/api.php"
    },
    {
      "id": "service:example_api",
      "title": "Example",
      "description": "Gives the probability for a taxon, given its
weight and location.",
      "url": "https://api.example.com"
    }
  ],
  "mediaElements": [
    {
      "id": "media:pointy_tail",
      "mediaElement": {
        "file": [
          {
            "url":
"https://github.com/WouterKoch/Clavis/raw/main/Keys/Images/pointy_100.
png",
            "width": 100,
            "height": 100,
            "license":
"https://creativecommons.org/licenses/by/4.0/",
            "creator": "person:wouterkoch"
          },
          {
            "url":
"https://github.com/WouterKoch/Clavis/raw/main/Keys/Images/pointy_250.
png",
            "width": 250,
            "height": 250,
            "license":
"https://creativecommons.org/licenses/by/4.0/",
            "creator": "person:wouterkoch"
          }
        ]
      }
    },
    {
      "id": "media:lobed_tail",
      "mediaElement": {
        "file": [
          {
            "url":
```

```
"https://github.com/WouterKoch/Clavis/raw/main/Keys/Images/lobed_100.p
ng",
            "width": 100,
            "height": 100,
            "license":
"https://creativecommons.org/licenses/by/4.0/",
            "creator": "person:wouterkoch"
          },
          {
            "url":
"https://github.com/WouterKoch/Clavis/raw/main/Keys/Images/lobed_250.p
ng",
            "width": 250,
            "height": 250,
            "license":
"https://creativecommons.org/licenses/by/4.0/",
            "creator": "person:wouterkoch"
          }
        ]
      }
    }
  ]
}
```

S3 Appendix. Titmice key example


```
{
   "$schema":
"https://raw.githubusercontent.com/WouterKoch/Clavis/main/Schema/Clavi
s.json",
   "title": "A key to titmice in Norway",
   "language": "en",
   "license": "https://creativecommons.org/licenses/by/4.0/",
   "creator": "person:wouterkoch",
   "lastModified": "2022-03-19 21:41:00",
   "identifier": "9be11d7e-c147-400a-899e-b3d5e4bcc6a1",
   "geography": {
      "name": "Norway",
      "polygon": [
         [
            [
               [
                  33.22265625,
                  69.56522590149099
               ],
               [
                  29.267578125,
                  71.15939141681443
               ],
               [
                  23.5546875,
                  71.28669893545877
               ],
               [
                  17.2265625,
                  69.97549253616164
               ],
               [
                  12.392578125,
                  68.366801093914
               ],
               [
                  11.2939453125,
                  65.91062334197893
               ],
               [
                  3.8671874999999996,
                  62.103882522897855
               ],
               [
                  4.7021484375,
                  58.516651799363785
```

```
                ],
                [
                   7.119140625,
                   57.70414723434193
                ],
                [
                   11.953125,
                   58.83649009392136
                ],
                [
                   13.3154296875,
                   61.39671887310411
                ],
                [
                   12.8759765625,
                   63.6267446447533
                ],
                [
                   14.94140625,
                   64.07219957867282
                ],
                [
                   15.2490234375,
                   66.05371622067922
                ],
                [
                   18.852539062499996,
                   68.02402198693447
                ],
                [
                   25.048828125,
                   68.51214331858073
                ],
                [
                   26.894531249999996,
                   69.54987728327795
                ],
                [
                   29.003906249999996,
                   68.8159271333607
                ],
                [
                   33.22265625,
                   69.56522590149099
                ]
             ]
          ]
       ]
```

```
   },
   "persons": [
      {
         "id": "person:wouterkoch",
         "name": "Wouter Koch"
      }
   ],
   "externalServices": [
      {
         "id": "service:nbic_taxa",
         "title": "NBIC taxonomy scientificNameId",
         "description": "To retrieve taxon information based on the
NBIC scientificNameId, e.g. through
https://www.artsdatabanken.no/api/Taxon/ByScientificNameId/4362",
         "provider": "Norwegian Biodiversity Information Centre",
         "url": "https://www.artsdatabanken.no/help"
      }
   ],
   "taxa": [
      {
         "id": "taxon:paridae",
         "scientificName": "Paridae",
         "rank": "family",
         "vernacularName": "titmice",
         "externalReference": {
            "serviceId": "service:nbic_taxa",
            "externalId": "4362"
         },
         "children": [
            {
               "id": "taxon:cyanistes",
               "scientificName": "Cyanistes",
               "rank": "genus",
               "externalReference": {
                  "serviceId": "service:nbic_taxa",
                  "externalId": "4364"
               },
               "children": [
                  {
                     "id": "taxon:cyanistes_caeruleus",
                     "scientificName": "Cyanistes caeruleus",
                     "vernacularName": "blue tit",
                     "rank": "species",
                     "externalReference": {
                        "serviceId": "service:nbic_taxa",
                        "externalId": "4365"
                     }
                  }
```

```
            ]
         },
         {
            "id": "taxon:lophophanes",
            "scientificName": "Lophophanes",
            "rank": "genus",
            "externalReference": {
               "serviceId": "service:nbic_taxa",
               "externalId": "4368"
            },
            "children": [
               {
                  "id": "taxon:lophophanes_cristatus",
                  "scientificName": "Lophophanes cristatus",
                  "vernacularName": "crested tit",
                  "rank": "species",
                  "externalReference": {
                     "serviceId": "service:nbic_taxa",
                     "externalId": "4369"
                  },
                  "geography": {
                     "polygon": [
                        [
                           [
                              [
                                 11.997070312499998,
                                 58.74540696858028
                              ],
                              [
                                 13.0517578125,
                                 59.977005492196
                              ],
                              [
                                 12.8759765625,
                                 63.31268278043484
                              ],
                              [
                                 19.2041015625,
                                 68.57644086491786
                              ],
                              [
                                 13.4912109375,
                                 68.78414378041504
                              ],
                              [
                                 9.7119140625,
                                 64.47279382008166
                              ],
```

```
                                [
                                    4.21875,
                                    62.2679226294176
                                ],
                                [
                                    3.69140625,
                                    59.40036514079251
                                ],
                                [
                                    6.723632812499999,
                                    57.70414723434193
                                ],
                                [
                                    11.997070312499998,
                                    58.74540696858028
                                ]
                            ]
                        ]
                    ]
                }
            }
        ]
    },
    {
        "id": "taxon:parus",
        "scientificName": "Parus",
        "rank": "genus",
        "externalReference": {
            "serviceId": "service:nbic_taxa",
            "externalId": "4363"
        },
        "children": [
            {
                "id": "taxon:parus_major",
                "scientificName": "Parus major",
                "vernacularName": "great tit",
                "rank": "species",
                "externalReference": {
                    "serviceId": "service:nbic_taxa",
                    "externalId": "4372"
                }
            }
        ]
    },
    {
        "id": "taxon:periparus",
        "scientificName": "Periparus",
        "rank": "genus",
```

```
            "externalReference": {
               "serviceId": "service:nbic_taxa",
               "externalId": "4374"
            },
            "children": [
               {
                  "id": "taxon:periparus_ater",
                  "scientificName": "Periparus ater",
                  "vernacularName": "coal tit",
                  "rank": "species",
                  "externalReference": {
                     "serviceId": "service:nbic_taxa",
                     "externalId": "4375"
                  }
               }
            ]
         },
         {
            "id": "taxon:poecile",
            "scientificName": "Poecile",
            "rank": "genus",
            "externalReference": {
               "serviceId": "service:nbic_taxa",
               "externalId": "4378"
            },
            "children": [
               {
                  "id": "taxon:poecile_palustris",
                  "scientificName": "Poecile palustris",
                  "vernacularName": "marsh tit",
                  "rank": "species",
                  "externalReference": {
                     "serviceId": "service:nbic_taxa",
                     "externalId": "4385"
                  },
                  "geography": {
                     "polygon": [
                        [
                           [
                              [
                                 11.997070312499998,
                                 58.74540696858028
                              ],
                              [
                                 13.0517578125,
                                 59.977005492196
                              ],
                              [
```

```
                                    12.8759765625,
                                    63.31268278043484
                                ],
                                [
                                    19.2041015625,
                                    68.57644086491786
                                ],
                                [
                                    13.4912109375,
                                    68.78414378041504
                                ],
                                [
                                    9.7119140625,
                                    64.47279382008166
                                ],
                                [
                                    4.21875,
                                    62.2679226294176
                                ],
                                [
                                    3.69140625,
                                    59.40036514079251
                                ],
                                [
                                    6.723632812499999,
                                    57.70414723434193
                                ],
                                [
                                    11.997070312499998,
                                    58.74540696858028
                                ]
                            ]
                        ]
                    ]
                }
            },
            {
                "id": "taxon:poecile_montanus",
                "scientificName": "Poecile montanus",
                "vernacularName": "willow tit",
                "rank": "species",
                "externalReference": {
                    "serviceId": "service:nbic_taxa",
                    "externalId": "4382"
                }
            },
            {
                "id": "taxon:poecile_cinctus",
```

```
                    "scientificName": "Poecile cinctus",
                    "vernacularName": "Siberian tit",
                    "rank": "species",
                    "externalReference": {
                       "serviceId": "service:nbic_taxa",
                       "externalId": "4379"
                    }
                 }
              ]
           }
        ]
     }
  ],
  "characters": [
     {
        "id": "character:head_top",
        "title": "Top of the head",
        "states": [
           {
              "id": "state:black_or_dark_grey",
              "title": "Black or dark grey"
           },
           {
              "id": "state:blue",
              "title": "Blue"
           },
           {
              "id": "state:speckled_crest",
              "title": "Speckled black and white, with a crest"
           }
        ]
     },
     {
        "id": "character:head_top",
        "title": "Top of the head",
        "states": [
           {
              "id": "state:black_or_dark_grey",
              "title": "Black or dark grey"
           },
           {
              "id": "state:brown",
              "title": "Brown"
           },
           {
              "id": "state:blue",
              "title": "Blue"
           },
```

```
            {
               "id": "state:speckled_crest",
               "title": "Speckled black and white, with a crest"
            }
         ]
      },
      {
         "id": "character:chest_color",
         "title": "Color of the chest",
         "states": [
            {
               "id": "state:yellow",
               "title": "Yellow"
            },
            {
               "id": "state:grey_brown",
               "title": "Grey to brown"
            }
         ]
      },
      {
         "id": "character:wing_bar",
         "title": "White bar on the wing",
         "states": [
            {
               "id": "state:wing_bar",
               "title": "Present"
            },
            {
               "id": "state:no_wing_bar",
               "title": "Absent"
            }
         ]
      },
      {
         "id": "character:black_of_cheek",
         "title": "Color of the cheek at the back",
         "states": [
            {
               "id": "state:white_cheek_back",
               "title": "Entire cheek white"
            },
            {
               "id": "state:brown_cheek_back",
               "title": "Sullied brown"
            }
         ]
      },
```

```
      {
         "id": "character:wing_secondaries_color",
         "title": "Color of secondary wing feathers",
         "states": [
            {
               "id": "state:secondaries_pale",
               "title": "Paler than rest of the wing"
            },
            {
               "id": "state:secondaries_not_pale",
               "title": "No clear different from rest of the wing"
            }
         ]
      }
   ],
   "statements": [
      {
         "id": "statement:crested_tit_head",
         "taxon": "taxon:lophophanes_cristatus",
         "character": "character:head_top",
         "value": "state:speckled_crest",
         "frequency": 1
      },
      {
         "id": "statement:blue_tit_head",
         "taxon": "taxon:cyanistes_caeruleus",
         "character": "character:head_top",
         "value": "state:blue",
         "frequency": 1
      },
      {
         "id": "statement:poecile_palustris_head",
         "taxon": "taxon:poecile_palustris",
         "character": "character:head_top",
         "value": "state:black_or_dark_grey",
         "frequency": 1
      },
      {
         "id": "statement:poecile_montanus_head",
         "taxon": "taxon:poecile_montanus",
         "character": "character:head_top",
         "value": "state:black_or_dark_grey",
         "frequency": 1
      },
      {
         "id": "statement:poecile_cinctus_head",
         "taxon": "taxon:poecile_cinctus",
         "character": "character:head_top",
```

```
            "value": "state:brown",
            "frequency": 1
        },
        {
            "id": "statement:great_tit_head",
            "taxon": "taxon:parus_major",
            "character": "character:head_top",
            "value": "state:black_or_dark_grey",
            "frequency": 1
        },
        {
            "id": "statement:coal_tit_head",
            "taxon": "taxon:periparus_ater",
            "character": "character:head_top",
            "value": "state:black_or_dark_grey",
            "frequency": 1
        },
        {
            "id": "statement:crested_tit_chest",
            "taxon": "taxon:lophophanes_cristatus",
            "character": "character:chest_color",
            "value": "state:grey_brown",
            "frequency": 1
        },
        {
            "id": "statement:blue_tit_chest",
            "taxon": "taxon:cyanistes_caeruleus",
            "character": "character:chest_color",
            "value": "state:yellow",
            "frequency": 1
        },
        {
            "id": "statement:poecile_chest",
            "taxon": "taxon:poecile",
            "character": "character:chest_color",
            "value": "state:grey_brown",
            "frequency": 1
        },
        {
            "id": "statement:great_tit_chest",
            "taxon": "taxon:parus_major",
            "character": "character:chest_color",
            "value": "state:yellow",
            "frequency": 1
        },
        {
            "id": "statement:coal_tit_chest",
            "taxon": "taxon:periparus_ater",
```

```
            "character": "character:chest_color",
            "value": "state:grey_brown",
            "frequency": 1
        },
        {
            "id": "statement:crested_tit_bar",
            "taxon": "taxon:lophophanes_cristatus",
            "character": "character:wing_bar",
            "value": "state:no_wing_bar",
            "frequency": 1
        },
        {
            "id": "statement:blue_tit_bar",
            "taxon": "taxon:cyanistes_caeruleus",
            "character": "character:wing_bar",
            "value": "state:wing_bar",
            "frequency": 1
        },
        {
            "id": "statement:poecile_bar",
            "taxon": "taxon:poecile",
            "character": "character:wing_bar",
            "value": "state:no_wing_bar",
            "frequency": 1
        },
        {
            "id": "statement:great_tit_bar",
            "taxon": "taxon:parus_major",
            "character": "character:wing_bar",
            "value": "state:wing_bar",
            "frequency": 1
        },
        {
            "id": "statement:coal_tit_bar",
            "taxon": "taxon:periparus_ater",
            "character": "character:wing_bar",
            "value": "state:wing_bar",
            "frequency": 1
        },
        {
            "id": "statement:poecile_palustris_cheek_back",
            "taxon": "taxon:poecile_palustris",
            "character": "character:black_of_cheek",
            "value": "state:brown_cheek_back",
            "frequency": 1
        },
        {
            "id": "statement:poecile_montanus_cheek_back",
```

```
            "taxon": "taxon:poecile_montanus",
            "character": "character:black_of_cheek",
            "value": "state:white_cheek_back",
            "frequency": 1
        },
        {
            "id": "statement:poecile_palustris_wing_secondaries_color",
            "taxon": "taxon:poecile_palustris",
            "character": "character:wing_secondaries_color",
            "value": "state:secondaries_not_pale",
            "frequency": 1
        },
        {
            "id": "statement:poecile_montanus_wing_secondaries_color",
            "taxon": "taxon:poecile_montanus",
            "character": "character:wing_secondaries_color",
            "value": "state:secondaries_pale",
            "frequency": 1
        }
    ]
}
```

Doctoral theses in Biology

Norwegian University of Science and Technology
Department of Biology

Year	Name	Degree	Title
1974	Tor-Henning Iversen	Dr. philos Botany	The roles of statholiths, auxin transport, and auxin metabolism in root gravitropism
1978	Tore Slagsvold	Dr. philos Zoology	Breeding events of birds in relation to spring temperature and environmental phenology
1978	Egil Sakshaug	Dr. philos Botany	The influence of environmental factors on the chemical composition of cultivated and natural populations of marine phytoplankton
1980	Arnfinn Langeland	Dr. philos Zoology	Interaction between fish and zooplankton populations and their effects on the material utilization in a freshwater lake
1980	Helge Reinertsen	Dr. philos Botany	The effect of lake fertilization on the dynamics and stability of a limnetic ecosystem with special reference to the phytoplankton
1982	Gunn Mari Olsen	Dr. scient Botany	Gravitropism in roots of *Pisum sativum* and *Arabidopsis thaliana*
1982	Dag Dolmen	Dr. philos Zoology	Life aspects of two sympartic species of newts (Triturus, Amphibia) in Norway, with special emphasis on their ecological niche segregation
1984	Eivin Røskaft	Dr. philos Zoology	Sociobiological studies of the rook *Corvus frugilegus*
1984	Anne Margrethe Cameron	Dr. scient Botany	Effects of alcohol inhalation on levels of circulating testosterone, follicle stimulating hormone and luteinzing hormone in male mature rats
1984	Asbjørn Magne Nilsen	Dr. scient Botany	Alveolar macrophages from expectorates – Biological monitoring of workers exposed to occupational air pollution. An evaluation of the AM-test
1985	Jarle Mork	Dr. philos Zoology	Biochemical genetic studies in fish
1985	John Solem	Dr. philos Zoology	Taxonomy, distribution and ecology of caddisflies (Trichoptera) in the Dovrefjell mountains
1985	Randi E. Reinertsen	Dr. philos Zoology	Energy strategies in the cold: Metabolic and thermoregulatory adaptations in small northern birds
1986	Bernt-Erik Sæther	Dr. philos Zoology	Ecological and evolutionary basis for variation in reproductive traits of some vertebrates: A comparative approach
1986	Torleif Holthe	Dr. philos Zoology	Evolution, systematics, nomenclature, and zoogeography in the polychaete orders Oweniimorpha and Terebellomorpha, with special reference to the Arctic and Scandinavian fauna
1987	Helene Lampe	Dr. scient Zoology	The function of bird song in mate attraction and territorial defence, and the importance of song repertoires
1987	Olav Hogstad	Dr. philos Zoology	Winter survival strategies of the Willow tit *Parus montanus*
1987	Jarle Inge Holten	Dr. philos Botany	Autecological investigations along a coust-inland transect at Nord-Møre, Central Norway
1987	Rita Kumar	Dr. scient Botany	Somaclonal variation in plants regenerated from cell cultures of *Nicotiana sanderae* and *Chrysanthemum morifolium*
1987	Bjørn Åge Tømmerås	Dr. scient Zoology	Olfaction in bark beetle communities: Interspecific interactions in regulation of colonization density, predator - prey relationship and host attraction

1988	Hans Christian Pedersen	Dr. philos Zoology	Reproductive behaviour in willow ptarmigan with special emphasis on territoriality and parental care
1988	Tor G. Heggberget	Dr. philos Zoology	Reproduction in Atlantic Salmon (*Salmo salar*): Aspects of spawning, incubation, early life history and population structure
1988	Marianne V. Nielsen	Dr. scient Zoology	The effects of selected environmental factors on carbon allocation/growth of larval and juvenile mussels (*Mytilus edulis*)
1988	Ole Kristian Berg	Dr. scient Zoology	The formation of landlocked Atlantic salmon (*Salmo salar* L.)
1989	John W. Jensen	Dr. philos Zoology	Crustacean plankton and fish during the first decade of the manmade Nesjø reservoir, with special emphasis on the effects of gill nets and salmonid growth
1989	Helga J. Vivås	Dr. scient Zoology	Theoretical models of activity pattern and optimal foraging: Predictions for the Moose *Alces alces*
1989	Reidar Andersen	Dr. scient Zoology	Interactions between a generalist herbivore, the moose Alces alces, and its winter food resources: a study of behavioural variation
1989	Kurt Ingar Draget	Dr. scient Botany	Alginate gel media for plant tissue culture
1990	Bengt Finstad	Dr. scient Zoology	Osmotic and ionic regulation in Atlantic salmon, rainbow trout and Arctic charr: Effect of temperature, salinity and season
1990	Hege Johannesen	Dr. scient Zoology	Respiration and temperature regulation in birds with special emphasis on the oxygen extraction by the lung
1990	Åse Krøkje	Dr. scient Botany	The mutagenic load from air pollution at two work-places with PAH-exposure measured with Ames Salmonella/microsome test
1990	Arne Johan Jensen	Dr. philos Zoology	Effects of water temperature on early life history, juvenile growth and prespawning migrations of Atlantic salmon (*Salmo salar*) and brown trout (*Salmo trutta*): A summary of studies in Norwegian streams
1990	Tor Jørgen Almaas	Dr. scient Zoology	Pheromone reception in moths: Response characteristics of olfactory receptor neurons to intra- and interspecific chemical cues
1990	Magne Husby	Dr. scient Zoology	Breeding strategies in birds: Experiments with the Magpie *Pica pica*
1991	Tor Kvam	Dr. scient Zoology	Population biology of the European lynx (*Lynx lynx*) in Norway
1991	Jan Henning L'Abêe Lund	Dr. philos Zoology	Reproductive biology in freshwater fish, brown trout *Salmo trutta* and roach *Rutilus rutilus* in particular
1991	Asbjørn Moen	Dr. philos Botany	The plant cover of the boreal uplands of Central Norway. I. Vegetation ecology of Sølendet nature reserve; haymaking fens and birch woodlands
1991	Else Marie Løbersli	Dr. scient Botany	Soil acidification and metal uptake in plants
1991	Trond Nordtug	Dr. scient Zoology	Reflectometric studies of photomechanical adaptation in superposition eyes of arthropods
1991	Thyra Solem	Dr. scient Botany	Age, origin and development of blanket mires in Central Norway
1991	Odd Terje Sandlund	Dr. philos Zoology	The dynamics of habitat use in the salmonid genera Coregonus and Salvelinus: Ontogenic niche shifts and polymorphism
1991	Nina Jonsson	Dr. philos Zoology	Aspects of migration and spawning in salmonids
1991	Atle Bones	Dr. scient Botany	Compartmentation and molecular properties of thioglucoside glucohydrolase (myrosinase)

1992	Torgrim Breiehagen	Dr. scient Zoology	Mating behaviour and evolutionary aspects of the breeding system of two bird species: the Temminck's stint and the Pied flycatcher
1992	Anne Kjersti Bakken	Dr. scient Botany	The influence of photoperiod on nitrate assimilation and nitrogen status in timothy (*Phleum pratense* L.)
1992	Tycho Anker-Nilssen	Dr. scient Zoology	Food supply as a determinant of reproduction and population development in Norwegian Puffins *Fratercula arctica*
1992	Bjørn Munro Jenssen	Dr. philos Zoology	Thermoregulation in aquatic birds in air and water: With special emphasis on the effects of crude oil, chemically treated oil and cleaning on the thermal balance of ducks
1992	Arne Vollan Aarset	Dr. philos Zoology	The ecophysiology of under-ice fauna: Osmotic regulation, low temperature tolerance and metabolism in polar crustaceans.
1993	Geir Slupphaug	Dr. scient Botany	Regulation and expression of uracil-DNA glycosylase and O6-methylguanine-DNA methyltransferase in mammalian cells
1993	Tor Fredrik Næsje	Dr. scient Zoology	Habitat shifts in coregonids.
1993	Yngvar Asbjørn Olsen	Dr. scient Zoology	Cortisol dynamics in Atlantic salmon, *Salmo salar* L.: Basal and stressor-induced variations in plasma levels and some secondary effects.
1993	Bård Pedersen	Dr. scient Botany	Theoretical studies of life history evolution in modular and clonal organisms
1993	Ole Petter Thangstad	Dr. scient Botany	Molecular studies of myrosinase in Brassicaceae
1993	Thrine L. M. Heggberget	Dr. scient Zoology	Reproductive strategy and feeding ecology of the Eurasian otter *Lutra lutra*.
1993	Kjetil Bevanger	Dr. scient Zoology	Avian interactions with utility structures, a biological approach.
1993	Kåre Haugan	Dr. scient Botany	Mutations in the replication control gene trfA of the broad host-range plasmid RK2
1994	Peder Fiske	Dr. scient Zoology	Sexual selection in the lekking great snipe (*Gallinago media*): Male mating success and female behaviour at the lek
1994	Kjell Inge Reitan	Dr. scient Botany	Nutritional effects of algae in first-feeding of marine fish larvae
1994	Nils Røv	Dr. scient Zoology	Breeding distribution, population status and regulation of breeding numbers in the northeast-Atlantic Great Cormorant *Phalacrocorax carbo carbo*
1994	Annette-Susanne Hoepfner	Dr. scient Botany	Tissue culture techniques in propagation and breeding of Red Raspberry (*Rubus idaeus* L.)
1994	Inga Elise Bruteig	Dr. scient Botany	Distribution, ecology and biomonitoring studies of epiphytic lichens on conifers
1994	Geir Johnsen	Dr. scient Botany	Light harvesting and utilization in marine phytoplankton: Species-specific and photoadaptive responses
1994	Morten Bakken	Dr. scient Zoology	Infanticidal behaviour and reproductive performance in relation to competition capacity among farmed silver fox vixens, *Vulpes vulpes*
1994	Arne Moksnes	Dr. philos Zoology	Host adaptations towards brood parasitism by the Cockoo
1994	Solveig Bakken	Dr. scient Botany	Growth and nitrogen status in the moss *Dicranum majus* Sm. as influenced by nitrogen supply
1994	Torbjørn Forseth	Dr. scient Zoology	Bioenergetics in ecological and life history studies of fishes.

1995	Olav Vadstein	Dr. philos Botany	The role of heterotrophic planktonic bacteria in the cycling of phosphorus in lakes: Phosphorus requirement, competitive ability and food web interactions
1995	Hanne Christensen	Dr. scient Zoology	Determinants of Otter *Lutra lutra* distribution in Norway: Effects of harvest, polychlorinated biphenyls (PCBs), human population density and competition with mink Mustela vision
1995	Svein Håkon Lorentsen	Dr. scient Zoology	Reproductive effort in the Antarctic Petrel *Thalassoica antarctica*; the effect of parental body size and condition
1995	Chris Jørgen Jensen	Dr. scient Zoology	The surface electromyographic (EMG) amplitude as an estimate of upper trapezius muscle activity
1995	Martha Kold Bakkevig	Dr. scient Zoology	The impact of clothing textiles and construction in a clothing system on thermoregulatory responses, sweat accumulation and heat transport
1995	Vidar Moen	Dr. scient Zoology	Distribution patterns and adaptations to light in newly introduced populations of *Mysis relicta* and constraints on Cladoceran and Char populations
1995	Hans Haavardsholm Blom	Dr. philos Botany	A revision of the *Schistidium apocarpum* complex in Norway and Sweden
1996	Jorun Skjærmo	Dr. scient Botany	Microbial ecology of early stages of cultivated marine fish; inpact fish-bacterial interactions on growth and survival of larvae
1996	Ola Ugedal	Dr. scient Zoology	Radiocesium turnover in freshwater fishes
1996	Ingibjørg Einarsdottir	Dr. scient Zoology	Production of Atlantic salmon (*Salmo salar*) and Arctic charr (*Salvelinus alpinus*): A study of some physiological and immunological responses to rearing routines
1996	Christina M. S. Pereira	Dr. scient Zoology	Glucose metabolism in salmonids: Dietary effects and hormonal regulation
1996	Jan Fredrik Børseth	Dr. scient Zoology	The sodium energy gradients in muscle cells of *Mytilus edulis* and the effects of organic xenobiotics
1996	Gunnar Henriksen	Dr. scient Zoology	Status of Grey seal *Halichoerus grypus* and Harbour seal *Phoca vitulina* in the Barents sea region
1997	Gunvor Øie	Dr. scient Botany	Eevalution of rotifer *Brachionus plicatilis* quality in early first feeding of turbot *Scophtalmus maximus* L. larvae
1997	Håkon Holien	Dr. scient Botany	Studies of lichens in spruce forest of Central Norway. Diversity, old growth species and the relationship to site and stand parameters
1997	Ole Reitan	Dr. scient Zoology	Responses of birds to habitat disturbance due to damming
1997	Jon Arne Grøttum	Dr. scient Zoology	Physiological effects of reduced water quality on fish in aquaculture
1997	Per Gustav Thingstad	Dr. scient Zoology	Birds as indicators for studying natural and human-induced variations in the environment, with special emphasis on the suitability of the Pied Flycatcher
1997	Torgeir Nygård	Dr. scient Zoology	Temporal and spatial trends of pollutants in birds in Norway: Birds of prey and Willow Grouse used as
1997	Signe Nybø	Dr. scient Zoology	Impacts of long-range transported air pollution on birds with particular reference to the dipper *Cinclus cinclus* in southern Norway
1997	Atle Wibe	Dr. scient Zoology	Identification of conifer volatiles detected by receptor neurons in the pine weevil (*Hylobius abietis*), analysed by gas chromatography linked to electrophysiology and to mass spectrometry
1997	Rolv Lundheim	Dr. scient Zoology	Adaptive and incidental biological ice nucleators

1997	Arild Magne Landa	Dr. scient Zoology	Wolverines in Scandinavia: ecology, sheep depredation and conservation
1997	Kåre Magne Nielsen	Dr. scient Botany	An evolution of possible horizontal gene transfer from plants to sail bacteria by studies of natural transformation in *Acinetobacter calcoacetius*
1997	Jarle Tufto	Dr. scient Zoology	Gene flow and genetic drift in geographically structured populations: Ecological, population genetic, and statistical models
1997	Trygve Hesthagen	Dr. philos Zoology	Population responses of Arctic charr (*Salvelinus alpinus* (L.)) and brown trout (*Salmo trutta* L.) to acidification in Norwegian inland waters
1997	Trygve Sigholt	Dr. philos Zoology	Control of Parr-smolt transformation and seawater tolerance in farmed Atlantic Salmon (*Salmo salar*) Effects of photoperiod, temperature, gradual seawater acclimation, NaCl and betaine in the diet
1997	Jan Østnes	Dr. scient Zoology	Cold sensation in adult and neonate birds
1998	Seethaledsumy Visvalingam	Dr. scient Botany	Influence of environmental factors on myrosinases and myrosinase-binding proteins
1998	Thor Harald Ringsby	Dr. scient Zoology	Variation in space and time: The biology of a House sparrow metapopulation
1998	Erling Johan Solberg	Dr. scient Zoology	Variation in population dynamics and life history in a Norwegian moose (*Alces alces*) population: consequences of harvesting in a variable environment
1998	Sigurd Mjøen Saastad	Dr. scient Botany	Species delimitation and phylogenetic relationships between the *Sphagnum recurvum* complex (Bryophyta): genetic variation and phenotypic plasticity
1998	Bjarte Mortensen	Dr. scient Botany	Metabolism of volatile organic chemicals (VOCs) in a head liver S9 vial equilibration system in vitro
1998	Gunnar Austrheim	Dr. scient Botany	Plant biodiversity and land use in subalpine grasslands. – A conservation biological approach
1998	Bente Gunnveig Berg	Dr. scient Zoology	Encoding of pheromone information in two related moth species
1999	Kristian Overskaug	Dr. scient Zoology	Behavioural and morphological characteristics in Northern Tawny Owls *Strix aluco*: An intra- and interspecific comparative approach
1999	Hans Kristen Stenøien	Dr. scient Botany	Genetic studies of evolutionary processes in various populations of nonvascular plants (mosses, liverworts and hornworts)
1999	Trond Arnesen	Dr. scient Botany	Vegetation dynamics following trampling and burning in the outlying haylands at Sølendet, Central Norway
1999	Ingvar Stenberg	Dr. scient Zoology	Habitat selection, reproduction and survival in the White-backed Woodpecker *Dendrocopos leucotos*
1999	Stein Olle Johansen	Dr. scient Botany	A study of driftwood dispersal to the Nordic Seas by dendrochronology and wood anatomical analysis
1999	Trina Falck Galloway	Dr. scient Zoology	Muscle development and growth in early life stages of the Atlantic cod (*Gadus morhua* L.) and Halibut (*Hippoglossus hippoglossus* L.)
1999	Marianne Giæver	Dr. scient Zoology	Population genetic studies in three gadoid species: blue whiting (*Micromisistius poutassou*), haddock (*Melanogrammus aeglefinus*) and cod (*Gadus morhua*) in the North-East Atlantic
1999	Hans Martin Hanslin	Dr. scient Botany	The impact of environmental conditions of density dependent performance in the boreal forest bryophytes *Dicranum majus*, *Hylocomium splendens*, *Plagiochila asplenigides*, *Ptilium crista-castrensis* and *Rhytidiadelphus lokeus*

1999	Ingrid Bysveen Mjølnerød	Dr. scient Zoology	Aspects of population genetics, behaviour and performance of wild and farmed Atlantic salmon (*Salmo salar*) revealed by molecular genetic techniques
1999	Else Berit Skagen	Dr. scient Botany	The early regeneration process in protoplasts from *Brassica napus* hypocotyls cultivated under various g-forces
1999	Stein-Are Sæther	Dr. philos Zoology	Mate choice, competition for mates, and conflicts of interest in the Lekking Great Snipe
1999	Katrine Wangen Rustad	Dr. scient Zoology	Modulation of glutamatergic neurotransmission related to cognitive dysfunctions and Alzheimer's disease
1999	Per Terje Smiseth	Dr. scient Zoology	Social evolution in monogamous families:
1999	Gunnbjørn Bremset	Dr. scient Zoology	Young Atlantic salmon (*Salmo salar* L.) and Brown trout (*Salmo trutta* L.) inhabiting the deep pool habitat, with special reference to their habitat use, habitat preferences and competitive interactions
1999	Frode Ødegaard	Dr. scient Zoology	Host specificity as a parameter in estimates of arthropod species richness
1999	Sonja Andersen	Dr. scient Zoology	Expressional and functional analyses of human, secretory phospholipase A2
2000	Ingrid Salvesen	Dr. scient Botany	Microbial ecology in early stages of marine fish: Development and evaluation of methods for microbial management in intensive larviculture
2000	Ingar Jostein Øien	Dr. scient Zoology	The Cuckoo (*Cuculus canorus*) and its host: adaptions and counteradaptions in a coevolutionary arms race
2000	Pavlos Makridis	Dr. scient Botany	Methods for the microbial control of live food used for the rearing of marine fish larvae
2000	Sigbjørn Stokke	Dr. scient Zoology	Sexual segregation in the African elephant (*Loxodonta africana*)
2000	Odd A. Gulseth	Dr. philos Zoology	Seawater tolerance, migratory behaviour and growth of Charr, (*Salvelinus alpinus*), with emphasis on the high Arctic Dieset charr on Spitsbergen, Svalbard
2000	Pål A. Olsvik	Dr. scient Zoology	Biochemical impacts of Cd, Cu and Zn on brown trout (*Salmo trutta*) in two mining-contaminated rivers in Central Norway
2000	Sigurd Einum	Dr. scient Zoology	Maternal effects in fish: Implications for the evolution of breeding time and egg size
2001	Jan Ove Evjemo	Dr. scient Zoology	Production and nutritional adaptation of the brine shrimp *Artemia sp.* as live food organism for larvae of marine cold water fish species
2001	Olga Hilmo	Dr. scient Botany	Lichen response to environmental changes in the managed boreal forest systems
2001	Ingebrigt Uglem	Dr. scient Zoology	Male dimorphism and reproductive biology in corkwing wrasse (*Symphodus melops* L.)
2001	Bård Gunnar Stokke	Dr. scient Zoology	Coevolutionary adaptations in avian brood parasites and their hosts
2002	Ronny Aanes	Dr. scient Zoology	Spatio-temporal dynamics in Svalbard reindeer (*Rangifer tarandus platyrhynchus*)
2002	Mariann Sandsund	Dr. scient Zoology	Exercise- and cold-induced asthma. Respiratory and thermoregulatory responses
2002	Dag-Inge Øien	Dr. scient Botany	Dynamics of plant communities and populations in boreal vegetation influenced by scything at Sølendet, Central Norway
2002	Frank Rosell	Dr. scient Zoology	The function of scent marking in beaver (*Castor fiber*)
2002	Janne Østvang	Dr. scient Botany	The Role and Regulation of Phospholipase A2 in Monocytes During Atherosclerosis Development

2002	Terje Thun	Dr. philos Biology	Dendrochronological constructions of Norwegian conifer chronologies providing dating of historical material
2002	Birgit Hafjeld Borgen	Dr. scient Biology	Functional analysis of plant idioblasts (Myrosin cells) and their role in defense, development and growth
2002	Bård Øyvind Solberg	Dr. scient Biology	Effects of climatic change on the growth of dominating tree species along major environmental gradients
2002	Per Winge	Dr. scient Biology	The evolution of small GTP binding proteins in cellular organisms. Studies of RAC GTPases in *Arabidopsis thaliana* and the Ral GTPase from *Drosophila melanogaster*
2002	Henrik Jensen	Dr. scient Biology	Causes and consequences of individual variation in fitness-related traits in house sparrows
2003	Jens Rohloff	Dr. philos Biology	Cultivation of herbs and medicinal plants in Norway – Essential oil production and quality control
2003	Åsa Maria O. Espmark Wibe	Dr. scient Biology	Behavioural effects of environmental pollution in threespine stickleback *Gasterosteus aculeatur* L.
2003	Dagmar Hagen	Dr. scient Biology	Assisted recovery of disturbed arctic and alpine vegetation – an integrated approach
2003	Bjørn Dahle	Dr. scient Biology	Reproductive strategies in Scandinavian brown bears
2003	Cyril Lebogang Taolo	Dr. scient Biology	Population ecology, seasonal movement and habitat use of the African buffalo (*Syncerus caffer*) in Chobe National Park, Botswana
2003	Marit Stranden	Dr. scient Biology	Olfactory receptor neurones specified for the same odorants in three related Heliothine species (*Helicoverpa armigera*, *Helicoverpa assulta* and *Heliothis virescens*)
2003	Kristian Hassel	Dr. scient Biology	Life history characteristics and genetic variation in an expanding species, *Pogonatum dentatum*
2003	David Alexander Rae	Dr. scient Biology	Plant- and invertebrate-community responses to species interaction and microclimatic gradients in alpine and Artic environments
2003	Åsa A Borg	Dr. scient Biology	Sex roles and reproductive behaviour in gobies and guppies: a female perspective
2003	Eldar Åsgard Bendiksen	Dr. scient Biology	Environmental effects on lipid nutrition of farmed Atlantic salmon (*Salmo salar* L.) parr and smolt
2004	Torkild Bakken	Dr. scient Biology	A revision of Nereidinae (Polychaeta, Nereididae)
2004	Ingar Pareliussen	Dr. scient Biology	Natural and Experimental Tree Establishment in a Fragmented Forest, Ambohitantely Forest Reserve, Madagascar
2004	Tore Brembu	Dr. scient Biology	Genetic, molecular and functional studies of RAC GTPases and the WAVE-like regulatory protein complex in *Arabidopsis thaliana*
2004	Liv S. Nilsen	Dr. scient Biology	Coastal heath vegetation on central Norway; recent past, present state and future possibilities
2004	Hanne T. Skiri	Dr. scient Biology	Olfactory coding and olfactory learning of plant odours in heliothine moths. An anatomical, physiological and behavioural study of three related species (*Heliothis virescens*, *Helicoverpa armigera* and *Helicoverpa assulta*)
2004	Lene Østby	Dr. scient Biology	Cytochrome P4501A (CYP1A) induction and DNA adducts as biomarkers for organic pollution in the natural environment
2004	Emmanuel J. Gerreta	Dr. philos Biology	The Importance of Water Quality and Quantity in the Tropical Ecosystems, Tanzania
2004	Linda Dalen	Dr. scient Biology	Dynamics of Mountain Birch Treelines in the Scandes Mountain Chain, and Effects of Climate Warming

2004	Lisbeth Mehli	Dr. scient Biology	Polygalacturonase-inhibiting protein (PGIP) in cultivated strawberry (*Fragaria x ananassa*): characterisation and induction of the gene following fruit infection by *Botrytis cinerea*
2004	Børge Moe	Dr. scient Biology	Energy-Allocation in Avian Nestlings Facing Short-Term Food Shortage
2005	Matilde Skogen Chauton	Dr. scient Biology	Metabolic profiling and species discrimination from High-Resolution Magic Angle Spinning NMR analysis of whole-cell samples
2005	Sten Karlsson	Dr. scient Biology	Dynamics of Genetic Polymorphisms
2005	Terje Bongard	Dr. scient Biology	Life History strategies, mate choice, and parental investment among Norwegians over a 300-year period
2005	Tonette Røstelien	PhD Biology	Functional characterisation of olfactory receptor neurone types in heliothine moths
2005	Erlend Kristiansen	Dr. scient Biology	Studies on antifreeze proteins
2005	Eugen G. Sørmo	Dr. scient Biology	Organochlorine pollutants in grey seal (*Halichoerus grypus*) pups and their impact on plasma thyroid hormone and vitamin A concentrations
2005	Christian Westad	Dr. scient Biology	Motor control of the upper trapezius
2005	Lasse Mork Olsen	PhD Biology	Interactions between marine osmo- and phagotrophs in different physicochemical environments
2005	Åslaug Viken	PhD Biology	Implications of mate choice for the management of small populations
2005	Ariaya Hymete Sahle Dingle	PhD Biology	Investigation of the biological activities and chemical constituents of selected Echinops spp. growing in Ethiopia
2005	Anders Gravbrøt Finstad	PhD Biology	Salmonid fishes in a changing climate: The winter challenge
2005	Shimane Washington Makabu	PhD Biology	Interactions between woody plants, elephants and other browsers in the Chobe Riverfront, Botswana
2005	Kjartan Østbye	Dr. scient Biology	The European whitefish *Coregonus lavaretus* (L.) species complex: historical contingency and adaptive radiation
2006	Kari Mette Murvoll	PhD Biology	Levels and effects of persistent organic pollutans (POPs) in seabirds, Retinoids and α-tocopherol – potential biomakers of POPs in birds?
2006	Ivar Herfindal	Dr. scient Biology	Life history consequences of environmental variation along ecological gradients in northern ungulates
2006	Nils Egil Tokle	PhD Biology	Are the ubiquitous marine copepods limited by food or predation? Experimental and field-based studies with main focus on *Calanus finmarchicus*
2006	Jan Ove Gjershaug	Dr. philos Biology	Taxonomy and conservation status of some booted eagles in south-east Asia
2006	Jon Kristian Skei	Dr. scient Biology	Conservation biology and acidification problems in the breeding habitat of amphibians in Norway
2006	Johanna Järnegren	PhD Biology	*Acesta oophaga* and *Acesta excavata* – a study of hidden biodiversity
2006	Bjørn Henrik Hansen	PhD Biology	Metal-mediated oxidative stress responses in brown trout (*Salmo trutta*) from mining contaminated rivers in Central Norway
2006	Vidar Grøtan	PhD Biology	Temporal and spatial effects of climate fluctuations on population dynamics of vertebrates
2006	Jafari R Kideghesho	PhD Biology	Wildlife conservation and local land use conflicts in Western Serengeti Corridor, Tanzania
2006	Anna Maria Billing	PhD Biology	Reproductive decisions in the sex role reversed pipefish *Syngnathus typhle*: when and how to invest in reproduction

2006	Henrik Pärn	PhD Biology	Female ornaments and reproductive biology in the bluethroat
2006	Anders J. Fjellheim	PhD Biology	Selection and administration of probiotic bacteria to marine fish larvae
2006	P. Andreas Svensson	PhD Biology	Female coloration, egg carotenoids and reproductive success: gobies as a model system
2007	Sindre A. Pedersen	PhD Biology	Metal binding proteins and antifreeze proteins in the beetle *Tenebrio molitor* - a study on possible competition for the semi-essential amino acid cysteine
2007	Kasper Hancke	PhD Biology	Photosynthetic responses as a function of light and temperature: Field and laboratory studies on marine microalgae
2007	Tomas Holmern	PhD Biology	Bushmeat hunting in the western Serengeti: Implications for community-based conservation
2007	Kari Jørgensen	PhD Biology	Functional tracing of gustatory receptor neurons in the CNS and chemosensory learning in the moth *Heliothis virescens*
2007	Stig Ulland	PhD Biology	Functional Characterisation of Olfactory Receptor Neurons in the Cabbage Moth, (*Mamestra brassicae* L.) (Lepidoptera, Noctuidae). Gas Chromatography Linked to Single Cell Recordings and Mass Spectrometry
2007	Snorre Henriksen	PhD Biology	Spatial and temporal variation in herbivore resources at northern latitudes
2007	Roelof Frans May	PhD Biology	Spatial Ecology of Wolverines in Scandinavia
2007	Vedasto Gabriel Ndibalema	PhD Biology	Demographic variation, distribution and habitat use between wildebeest sub-populations in the Serengeti National Park, Tanzania
2007	Julius William Nyahongo	PhD Biology	Depredation of Livestock by wild Carnivores and Illegal Utilization of Natural Resources by Humans in the Western Serengeti, Tanzania
2007	Shombe Ntaraluka Hassan	PhD Biology	Effects of fire on large herbivores and their forage resources in Serengeti, Tanzania
2007	Per-Arvid Wold	PhD Biology	Functional development and response to dietary treatment in larval Atlantic cod (*Gadus morhua* L.) Focus on formulated diets and early weaning
2007	Anne Skjetne Mortensen	PhD Biology	Toxicogenomics of Aryl Hydrocarbon- and Estrogen Receptor Interactions in Fish: Mechanisms and Profiling of Gene Expression Patterns in Chemical Mixture Exposure Scenarios
2008	Brage Bremset Hansen	PhD Biology	The Svalbard reindeer (*Rangifer tarandus platyrhynchus*) and its food base: plant-herbivore interactions in a high-arctic ecosystem
2008	Jiska van Dijk	PhD Biology	Wolverine foraging strategies in a multiple-use landscape
2008	Flora John Magige	PhD Biology	The ecology and behaviour of the Masai Ostrich (*Struthio camelus massaicus*) in the Serengeti Ecosystem, Tanzania
2008	Bernt Rønning	PhD Biology	Sources of inter- and intra-individual variation in basal metabolic rate in the zebra finch, *Taeniopygia guttata*
2008	Sølvi Wehn	PhD Biology	Biodiversity dynamics in semi-natural mountain landscapes - A study of consequences of changed agricultural practices in Eastern Jotunheimen
2008	Trond Moxness Kortner	PhD Biology	The Role of Androgens on previtellogenic oocyte growth in Atlantic cod (*Gadus morhua*): Identification and patterns of differentially expressed genes in relation to Stereological Evaluations
2008	Katarina Mariann Jørgensen	Dr. scient Biology	The role of platelet activating factor in activation of growth arrested keratinocytes and re-epithelialisation

2008	Tommy Jørstad	PhD Biology	Statistical Modelling of Gene Expression Data
2008	Anna Kusnierczyk	PhD Biology	Arabidopsis thaliana Responses to Aphid Infestation
2008	Jussi Evertsen	PhD Biology	Herbivore sacoglossans with photosynthetic chloroplasts
2008	John Eilif Hermansen	PhD Biology	Mediating ecological interests between locals and globals by means of indicators. A study attributed to the asymmetry between stakeholders of tropical forest at Mt. Kilimanjaro, Tanzania
2008	Ragnhild Lyngved	PhD Biology	Somatic embryogenesis in *Cyclamen persicum*. Biological investigations and educational aspects of cloning
2008	Line Elisabeth Sundt-Hansen	PhD Biology	Cost of rapid growth in salmonid fishes
2008	Line Johansen	PhD Biology	Exploring factors underlying fluctuations in white clover populations – clonal growth, population structure and spatial distribution
2009	Astrid Jullumstrø Feuerherm	PhD Biology	Elucidation of molecular mechanisms for pro-inflammatory phospholipase A2 in chronic disease
2009	Pål Kvello	PhD Biology	Neurons forming the network involved in gustatory coding and learning in the moth *Heliothis virescens*: Physiological and morphological characterisation, and integration into a standard brain atlas
2009	Trygve Devold Kjellsen	PhD Biology	Extreme Frost Tolerance in Boreal Conifers
2009	Johan Reinert Vikan	PhD Biology	Coevolutionary interactions between common cuckoos *Cuculus canorus* and Fringilla finches
2009	Zsolt Volent	PhD Biology	Remote sensing of marine environment: Applied surveillance with focus on optical properties of phytoplankton, coloured organic matter and suspended matter
2009	Lester Rocha	PhD Biology	Functional responses of perennial grasses to simulated grazing and resource availability
2009	Dennis Ikanda	PhD Biology	Dimensions of a Human-lion conflict: Ecology of human predation and persecution of African lions (*Panthera leo*) in Tanzania
2010	Huy Quang Nguyen	PhD Biology	Egg characteristics and development of larval digestive function of cobia (*Rachycentron canadum*) in response to dietary treatments - Focus on formulated diets
2010	Eli Kvingedal	PhD Biology	Intraspecific competition in stream salmonids: the impact of environment and phenotype
2010	Sverre Lundemo	PhD Biology	Molecular studies of genetic structuring and demography in Arabidopsis from Northern Europe
2010	Iddi Mihijai Mfunda	PhD Biology	Wildlife Conservation and People's livelihoods: Lessons Learnt and Considerations for Improvements. The Case of Serengeti Ecosystem, Tanzania
2010	Anton Tinchov Antonov	PhD Biology	Why do cuckoos lay strong-shelled eggs? Tests of the puncture resistance hypothesis
2010	Anders Lyngstad	PhD Biology	Population Ecology of *Eriophorum latifolium*, a Clonal Species in Rich Fen Vegetation
2010	Hilde Færevik	PhD Biology	Impact of protective clothing on thermal and cognitive responses
2010	Ingerid Brænne Arbo	PhD Medical technology	Nutritional lifestyle changes – effects of dietary carbohydrate restriction in healthy obese and overweight humans
2010	Yngvild Vindenes	PhD Biology	Stochastic modeling of finite populations with individual heterogeneity in vital parameters
2010	Hans-Richard Brattbakk	PhD Medical technology	The effect of macronutrient composition, insulin stimulation, and genetic variation on leukocyte gene expression and possible health benefits

2011	Geir Hysing Bolstad	PhD Biology	Evolution of Signals: Genetic Architecture, Natural Selection and Adaptive Accuracy
2011	Karen de Jong	PhD Biology	Operational sex ratio and reproductive behaviour in the two-spotted goby (*Gobiusculus flavescens*)
2011	Ann-Iren Kittang	PhD Biology	*Arabidopsis thaliana* L. adaptation mechanisms to microgravity through the EMCS MULTIGEN-2 experiment on the ISS: The science of space experiment integration and adaptation to simulated microgravity
2011	Aline Magdalena Lee	PhD Biology	Stochastic modeling of mating systems and their effect on population dynamics and genetics
2011	Christopher Gravningen Sørmo	PhD Biology	Rho GTPases in Plants: Structural analysis of ROP GTPases; genetic and functional studies of MIRO GTPases in *Arabidopsis thaliana*
2011	Grethe Robertsen	PhD Biology	Relative performance of salmonid phenotypes across environments and competitive intensities
2011	Line-Kristin Larsen	PhD Biology	Life-history trait dynamics in experimental populations of guppy (*Poecilia reticulata*): the role of breeding regime and captive environment
2011	Maxim A. K. Teichert	PhD Biology	Regulation in Atlantic salmon (*Salmo salar*): The interaction between habitat and density
2011	Torunn Beate Hancke	PhD Biology	Use of Pulse Amplitude Modulated (PAM) Fluorescence and Bio-optics for Assessing Microalgal Photosynthesis and Physiology
2011	Sajeda Begum	PhD Biology	Brood Parasitism in Asian Cuckoos: Different Aspects of Interactions between Cuckoos and their Hosts in Bangladesh
2011	Kari J. K. Attramadal	PhD Biology	Water treatment as an approach to increase microbial control in the culture of cold water marine larvae
2011	Camilla Kalvatn Egset	PhD Biology	The Evolvability of Static Allometry: A Case Study
2011	AHM Raihan Sarker	PhD Biology	Conflict over the conservation of the Asian elephant (*Elephas maximus*) in Bangladesh
2011	Gro Dehli Villanger	PhD Biology	Effects of complex organohalogen contaminant mixtures on thyroid hormone homeostasis in selected arctic marine mammals
2011	Kari Bjørneraas	PhD Biology	Spatiotemporal variation in resource utilisation by a large herbivore, the moose
2011	John Odden	PhD Biology	The ecology of a conflict: Eurasian lynx depredation on domestic sheep
2011	Simen Pedersen	PhD Biology	Effects of native and introduced cervids on small mammals and birds
2011	Mohsen Falahati-Anbaran	PhD Biology	Evolutionary consequences of seed banks and seed dispersal in Arabidopsis
2012	Jakob Hønborg Hansen	PhD Biology	Shift work in the offshore vessel fleet: circadian rhythms and cognitive performance
2012	Elin Noreen	PhD Biology	Consequences of diet quality and age on life-history traits in a small passerine bird
2012	Irja Ida Ratikainen	PhD Biology	Foraging in a variable world: adaptations to stochasticity
2012	Aleksander Handå	PhD Biology	Cultivation of mussels (*Mytilus edulis*): Feed requirements, storage and integration with salmon (*Salmo salar*) farming
2012	Morten Kraabøl	PhD Biology	Reproductive and migratory challenges inflicted on migrant brown trout (*Salmo trutta* L.) in a heavily modified river
2012	Jisca Huisman	PhD Biology	Gene flow and natural selection in Atlantic salmon
2012	Maria Bergvik	PhD Biology	Lipid and astaxanthin contents and biochemical post-harvest stability in *Calanus finmarchicus*

2012	Bjarte Bye Løfaldli	PhD Biology	Functional and morphological characterization of central olfactory neurons in the model insect *Heliothis virescens*.
2012	Karen Marie Hammer	PhD Biology	Acid-base regulation and metabolite responses in shallow- and deep-living marine invertebrates during environmental hypercapnia
2012	Øystein Nordrum Wiggen	PhD Biology	Optimal performance in the cold
2012	Robert Dominikus Fyumagwa	Dr. Philos Biology	Anthropogenic and natural influence on disease prevalence at the human –livestock-wildlife interface in the Serengeti ecosystem, Tanzania
2012	Jenny Bytingsvik	PhD Biology	Organohalogenated contaminants (OHCs) in polar bear mother-cub pairs from Svalbard, Norway. Maternal transfer, exposure assessment and thyroid hormone disruptive effects in polar bear cubs
2012	Christer Moe Rolandsen	PhD Biology	The ecological significance of space use and movement patterns of moose in a variable environment
2012	Erlend Kjeldsberg Hovland	PhD Biology	Bio-optics and Ecology in *Emiliania huxleyi* Blooms: Field and Remote Sensing Studies in Norwegian Waters
2012	Lise Cats Myhre	PhD Biology	Effects of the social and physical environment on mating behaviour in a marine fish
2012	Tonje Aronsen	PhD Biology	Demographic, environmental and evolutionary aspects of sexual selection
2012	Bin Liu	PhD Biology	Molecular genetic investigation of cell separation and cell death regulation in *Arabidopsis thaliana*
2013	Jørgen Rosvold	PhD Biology	Ungulates in a dynamic and increasingly human dominated landscape – A millennia-scale perspective
2013	Pankaj Barah	PhD Biology	Integrated Systems Approaches to Study Plant Stress Responses
2013	Marit Linnerud	PhD Biology	Patterns in spatial and temporal variation in population abundances of vertebrates
2013	Xinxin Wang	PhD Biology	Integrated multi-trophic aquaculture driven by nutrient wastes released from Atlantic salmon (*Salmo salar*) farming
2013	Ingrid Ertshus Mathisen	PhD Biology	Structure, dynamics, and regeneration capacity at the sub-arctic forest-tundra ecotone of northern Norway and Kola Peninsula, NW Russia
2013	Anders Foldvik	PhD Biology	Spatial distributions and productivity in salmonid populations
2013	Anna Marie Holand	PhD Biology	Statistical methods for estimating intra- and inter-population variation in genetic diversity
2013	Anna Solvang Båtnes	PhD Biology	Light in the dark – the role of irradiance in the high Arctic marine ecosystem during polar night
2013	Sebastian Wacker	PhD Biology	The dynamics of sexual selection: effects of OSR, density and resource competition in a fish
2013	Cecilie Miljeteig	PhD Biology	Phototaxis in *Calanus finmarchicus* – light sensitivity and the influence of energy reserves and oil exposure
2013	Ane Kjersti Vie	PhD Biology	Molecular and functional characterisation of the IDA family of signalling peptides in *Arabidopsis thaliana*
2013	Marianne Nymark	PhD Biology	Light responses in the marine diatom *Phaeodactylum tricornutum*
2014	Jannik Schultner	PhD Biology	Resource Allocation under Stress - Mechanisms and Strategies in a Long-Lived Bird
2014	Craig Ryan Jackson	PhD Biology	Factors influencing African wild dog (*Lycaon pictus*) habitat selection and ranging behaviour: conservation and management implications

2014	Aravind Venkatesan	PhD Biology	Application of Semantic Web Technology to establish knowledge management and discovery in the Life Sciences
2014	Kristin Collier Valle	PhD Biology	Photoacclimation mechanisms and light responses in marine micro- and macroalgae
2014	Michael Puffer	PhD Biology	Effects of rapidly fluctuating water levels on juvenile Atlantic salmon (*Salmo salar* L.)
2014	Gundula S. Bartzke	PhD Biology	Effects of power lines on moose (*Alces alces*) habitat selection, movements and feeding activity
2014	Eirin Marie Bjørkvoll	PhD Biology	Life-history variation and stochastic population dynamics in vertebrates
2014	Håkon Holand	PhD Biology	The parasite *Syngamus trachea* in a metapopulation of house sparrows
2014	Randi Magnus Sommerfelt	PhD Biology	Molecular mechanisms of inflammation – a central role for cytosolic phospholiphase A2
2014	Espen Lie Dahl	PhD Biology	Population demographics in white-tailed eagle at an on-shore wind farm area in coastal Norway
2014	Anders Øverby	PhD Biology	Functional analysis of the action of plant isothiocyanates: cellular mechanisms and in vivo role in plants, and anticancer activity
2014	Kamal Prasad Acharya	PhD Biology	Invasive species: Genetics, characteristics and trait variation along a latitudinal gradient.
2014	Ida Beathe Øverjordet	PhD Biology	Element accumulation and oxidative stress variables in Arctic pelagic food chains: Calanus, little auks (*Alle alle*) and black-legged kittiwakes (*Rissa tridactyla*)
2014	Kristin Møller Gabrielsen	PhD Biology	Target tissue toxicity of the thyroid hormone system in two species of arctic mammals carrying high loads of organohalogen contaminants
2015	Gine Roll Skjervø	Dr. philos Biology	Testing behavioral ecology models with historical individual-based human demographic data from Norway
2015	Nils Erik Gustaf Forsberg	PhD Biology	Spatial and Temporal Genetic Structure in Landrace Cereals
2015	Leila Alipanah	PhD Biology	Integrated analyses of nitrogen and phosphorus deprivation in the diatoms *Phaeodactylum tricornutum* and *Seminavis robusta*
2015	Javad Najafi	PhD Biology	Molecular investigation of signaling components in sugar sensing and defense in *Arabidopsis thaliana*
2015	Bjørnar Sporsheim	PhD Biology	Quantitative confocal laser scanning microscopy: optimization of in vivo and in vitro analysis of intracellular transport
2015	Magni Olsen Kyrkjeeide	PhD Biology	Genetic variation and structure in peatmosses (Sphagnum)
2015	Keshuai Li	PhD Biology	Phospholipids in Atlantic cod (*Gadus morhua* L.) larvae rearing: Incorporation of DHA in live feed and larval phospholipids and the metabolic capabilities of larvae for the de novo synthesis
2015	Ingvild Fladvad Størdal	PhD Biology	The role of the copepod *Calanus finmarchicus* in affecting the fate of marine oil spills
2016	Thomas Kvalnes	PhD Biology	Evolution by natural selection in age-structured populations in fluctuating environments
2016	Øystein Leiknes	PhD Biology	The effect of nutrition on important life-history traits in the marine copepod *Calanus finmarchicus*
2016	Johan Henrik Hårdensson Berntsen	PhD Biology	Individual variation in survival: The effect of incubation temperature on the rate of physiological ageing in a small passerine bird
2016	Marianne Opsahl Olufsen	PhD Biology	Multiple environmental stressors: Biological interactions between parameters of climate change and perfluorinated alkyl substances in fish

2016	Rebekka Varne	PhD Biology	Tracing the fate of escaped cod (*Gadus morhua* L.) in a Norwegian fjord system
2016	Anette Antonsen Fenstad	PhD Biology	Pollutant Levels, Antioxidants and Potential Genotoxic Effects in Incubating Female Common Eiders (*Somateria mollissima*)
2016	Wilfred Njama Marealle	PhD Biology	Ecology, Behaviour and Conservation Status of Masai Giraffe (*Giraffa camelopardalis tippelskirchi*) in Tanzania
2016	Ingunn Nilssen	PhD Biology	Integrated Enviromental Mapping and Monitoring: A Methodological approach for end users.
2017	Konika Chawla	PhD Biology	Discovering, analysing and taking care of knowledge.
2017	Øystein Hjorthol Opedal	PhD Biology	The Evolution of Herkogamy: Pollinator Reliability, Natural Selection, and Trait Evolvability.
2017	Ane Marlene Myhre	PhD Biology	Effective size of density dependent populations in fluctuating environments
2017	Emmanuel Hosiana Masenga	PhD Biology	Behavioural Ecology of Free-ranging and Reintroduced African Wild Dog (*Lycaon pictus*) Packs in the Serengeti Ecosystem, Tanzania
2017	Xiaolong Lin	PhD Biology	Systematics and evolutionary history of Tanytarsus van der Wulp, 1874 (Diptera: Chironomidae)
2017	Emmanuel Clamsen Mmassy	PhD Biology	Ecology and Conservation Challenges of the Kori bustard in the Serengeti National Park
2017	Richard Daniel Lyamuya	PhD Biology	Depredation of Livestock by Wild Carnivores in the Eastern Serengeti Ecosystem, Tanzania
2017	Katrin Hoydal	PhD Biology	Levels and endocrine disruptive effects of legacy POPs and their metabolites in long-finned pilot whales of the Faroe Islands
2017	Berit Glomstad	PhD Biology	Adsorption of phenanthrene to carbon nanotubes and its influence on phenanthrene bioavailability/toxicity in aquatic organism
2017	Øystein Nordeide Kielland	PhD Biology	Sources of variation in metabolism of an aquatic ectotherm
2017	Narjes Yousefi	PhD Biology	Genetic divergence and speciation in northern peatmosses (Sphagnum)
2018	Signe Christensen-Dalgaard	PhD Biology	Drivers of seabird spatial ecology - implications for development of offshore wind-power in Norway
2018	Janos Urbancsok	PhD Biology	Endogenous biological effects induced by externally supplemented glucosinolate hydrolysis products (GHPs) on *Arabidopsis thaliana*
2018	Alice Mühlroth	PhD Biology	The influence of phosphate depletion on lipid metabolism of microalgae
2018	Franco Peniel Mbise	PhD Biology	Human-Carnivore Coexistence and Conflict in the Eastern Serengeti, Tanzania
2018	Stine Svalheim Markussen	PhD Biology	Causes and consequences of intersexual life history variation in a harvested herbivore population
2018	Mia Vedel Sørensen	PhD Biology	Carbon budget consequences of deciduous shrub expansion in alpine tundra ecosystems
2018	Hanna Maria Kauko	PhD Biology	Light response and acclimation of microalgae in a changing Arctic
2018	Erlend I. F. Fossen	PhD Biology	Trait evolvability: effects of thermal plasticity and genetic correlations among traits
2019	Peter Sjolte Ranke	PhD Biology	Demographic and genetic and consequences of dispersal in house sparrows
2019	Mathilde Le Moullec	PhD Biology	Spatiotemporal variation in abundance of key tundra species: from local heterogeneity to large-scale synchrony
2019	Endre Grüner Ofstad	PhD Biology	Causes and consequences of variation in resource use and social structure in ungulates
2019	Yang Jin	PhD Biology	Development of lipid metabolism in early life stage of Atlantic salmon (*Salmo salar*)

2019	Elena Albertsen	PhD Biology	Evolution of floral traits: from ecological contex to functional integration
2019	Mominul Islam Nahid	PhD Biology	Interaction between two Asian cuckoos and their hosts in Bangladesh
2019	Knut Jørgen Egelie	PhD Biology	Management of intellectual property in university-industry collaborations – public access to and control of knowledge
2019	Thomas Ray Haaland	PhD Biology	Adaptive responses to enviromental stochasticity on different evolutionary time-scales
2019	Kwaslema Malle Hariohay	PhD Biology	Human wildlife interactions in the Ruaha-Rungwa Ecosystem, Central Tanzania
2019	Mari Engvig Løseth	PhD Biology	Exposure and effects of emerging and legacy organic pollutants in white-tailed eagle (*Haliaeetis albicilla*) nestlings
2019	Joseph Mbyati Mukeka	PhD Biology	Human-Wildlife Conflicts and Compensation for Losses in Kenya: Dynamics, Characteristics and Correlates
2019	Helene Løvstrand Svarva	PhD Biology	Dendroclimatology in southern Norway: tree rings, demography and climate
2019	Nathalie Briels	PhD Biology	Exposure and effects of legacy and emerging organic pollutants in developing birds – Laboratory and field studies
2019	Anders L.Kolstad	PhD Biology	Moose browsing effects on boreal production forests – implications for ecosystems and human society
2019	Bart Peeters	PhD Biology	Population dynamics under climate change ad harvesting: Results from the high Arctic Svalbard reindeer
2019	Emma-Liina Marjakangas	PhD Biology	Understanding species interactions in the tropics: dynamics within and between trophic levels
2019	Alex Kojo Datsomor	PhD Biology	The molecular basis of long chain polyunsaturated fatty acid (LC-PUFA) biosynthesis in Atlantic salmon (*Salmo salar* L): In vivo functions, functional redundancy and transcriptional regulation of LC-PUFA biosynthetic enzymes
2020	Ingun Næve	PhD Biology	Development of non-invasive methods using ultrasound technology in monitoring of Atlantic salmon (*Salmo salar*) production and reproduction
2020	Rachael Morgan	PhD Biology	Physiological plasticity and evolution of thermal performance in zebrafish
2020	Mahsa Jalili	PhD Biology	Effects of different dietary ingredients on the immune responses and antioxidant status in Atlantic salmon (*Salmo salar* L.): possible nutriomics approaches
2020	Haiqing Wang	PhD Biology	Utilization of the polychaete *Hediste diversicolor* (O.F. Millier, 1776) in recycling waste nutrients from land-based fish farms for value adding applications'
2020	Louis Hunninck	PhD Biology	Physiological and behavioral adaptations of impala to anthropogenic disturbances in the Serengeti ecosystems
2020	Kate Layton-Matthews	PhD Biology	Demographic consequences of rapid climate change and density dependence in migratory Arctic geese
2020	Amit Kumar Sharma	PhD Biology	Genome editing of marine algae: Technology development and use of the CRISPR/Cas9 system for studies of light harvesting complexes and regulation of phosphate homeostasis
2020	Lars Rød-Eriksen	PhD Biology	Drivers of change in meso-carnivore distributions in a northern ecosystem
2020	Lone Sunniva Jevne	PhD Biology	Development and dispersal of salmon lice (*Lepeophtheirus salmonis* Krøyer, 1837) in commercial salmon farming localities

2020	Sindre Håvarstein Eldøy	PhD Biology	The influence of physiology, life history and environmental conditions on the marine migration patters of sea trout
2020	Vasundra Touré	PhD Biology	Improving the FAIRness of causal interactions in systems biology: data curation and standardisation to support systems modelling applications
2020	Silje Forbord	PhD Biology	Cultivation of *Saccharina latissima* (Phaeophyceae) in temperate marine waters; nitrogen uptake kinetics, growth characteristics and chemical composition
2020	Jørn Olav Løkken	PhD Biology	Change in vegetation composition and growth in the forest-tundra ecotone – effects of climate warming and herbivory
2020	Kristin Odden Nystuen	PhD Biology	Drivers of plant recruitment in alpine vegetation
2021	Sam Perrin	PhD Biology	Freshwater Fish Community Responses to Climate Change and Invasive Species
2021	Lara Veylit	PhD Biology	Causes and consequences of body growth variation in hunted wild boar populations
2021	Semona Issa	PhD Biology	Combined effects of environmental variation and pollution on zooplankton life history and population dynamics
2021	Monica Shilereyo	PhD Biology	Small Mammal Population Ecology and Ectoparasite Load: Assessing Impacts of Land Use and Rainfall Seasonality in the Serengeti Ecosystem, Tanzania
2021	Vanessa Bieker	PhD Biology	Using historical herbarium specimens to elucidate the evolutionary genomics of plant invasion
2021	Håkon Austad Langberg	PhD Biology	Fate and transport of forever chemicals in the aquatic environment: Partitioning and biotransformation of mixtures of Per- and Polyfluoroalkyl Substances (PFAS) from different point sources and resulting concentrations in biota
2021	Julie Renberg	PhD Biology	Muscular and metabolic load and manual function when working in the cold
2021	Olena Meleshko	PhD Biology	Gene flow and genome evolution on peatmosses (Sphagnum)
2021	Essa Ahsan Khan	PhD Biology	Systems toxicology approach for evaluating the effects of contaminants on fish ovarian development and reproductive endocrine physiology:A combination of field-, in vivo and ex vivo studies using Atlantic cod (*Gadus morhua*)
2021	Tanja Kofod Petersen	PhD Biology	Biodiversity dynamics in urban areas under changing land-uses
2021	Katariina Vuorinen	PhD Biology	When do ungulates override the climate? Defining the interplay of two key drivers of northern vegetation dynamics
2021	Archana Golla	PhD Biology	Impact of early life stress on behaviour and dorsal raphe serotonergic activity in zebrafish (*Danio rerio*)
2021	Aksel Alstad Mogstad	PhD Biology	Underwater Hyperspectral Imaging as a Tool for Benthic Habitat Mapping
2021	Randi Grønnstad	PhD Biology	Per- and polyfluoroalkyl substances (PFAS) in ski products: Environmental contamination, bioaccumulation and effects in rodents
2021	Gaspard Philis	PhD Biology	Life cycle assessment of sea lice treatments in Norwegian net pens with emphasis on the environmental tradeoffs of salmon aquaculture production systems
2021	Christoffer Høyvik Hilde	PhD Biology	Demographic buffering of vital rates in age-structured populations

2021	Halldis Ringvold	Dr.Philos	Studies on Echinodermata from the NE Atlantic Ocean - Spatial distribution and abundance of Asteroidea, including taxonomic and molecular studies on Crossaster and Henricia genera- Value-chain results, including test fishery, biology, market and nutritional analysis, on *Parastichopus tremulus* (Holothuroidea) from the Norwegian coast
2021	Elise Skottene	PhD Biology	Lipid metabolism and diapause timing in Calanus copepods. The impact of predation risk, food availability and oil exposure
2021	Michael Le Pepke	PhD Biology	The ecological and evolutionary role of telomere length in house sparrows
2022	Niklas Erik Johansson	Dr. Philos	On the taxonomy of Northern European Darwin wasps (Hymenoptera: Ichneumonidae).
2022	Jonatan Fredricson Marquez	PhD Biology	Understanding spatial and interspecific processes affecting population dynamics in a marine ecosystem.
2022	Anne Mehlhoop	PhD Biology	Evaluating mitigation measures to reduce negative impacts of infrastructure construction on vegetation and wildlife.
2022	Malene Østreng Nygård	PhD Biology	Integrative biosystematics and conservation genomics – holistic studies of two red-listed plants in Norway
2022	Martin René Ellegaard	PhD Biology	Human Population Genomics in Northern Europe in the Past 2000 years
2022	Gaute Kjærstad	PhD Biology	The eradication of invasive species using rotenone and its impact on freshwater macroinvertebrates
2022	Stefan Vriend	PhD Biology	On the roles of density dependence and environmental fluctuations in driving eco-evolutionary dynamics of hole-nesting passerines
2022	Zaw Min Thant	PhD Biology	Anthropogenic and Environmental factors driving the Human-Elephant Conflict in Myanmar
2022	Prashanna Guragain	PhD Biology	Population analysis and structure and RNA interference to understand salmon lice biology and a review of the principles of controlling infestation in aquaculture facilities.
2022	Ronja Wedegärtner	PhD Biology	Highways up the mountains? Trails as facilitators for redistribution of plant species in mountain areas

ISBN 978-82-326-6083-4 (printed ver.)
ISBN 978-82-326-6669-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology