A minimum data standard for wildlife disease studies

Tess Stevens[1], Ryan Zimmerman[1], Greg Albery[2], Daniel J. Becker[3], Rebekah C. Kading[4], Carl N. Keiser[5], Shashank Khandelwal[6], Stephanie Kramer-Schadt[7], Raphael Krut-Landau[6], Clifton McKee[8], Diego Montecino-Latorre[9], Zoe O'Donoghue[1], Sarah H. Olson[9], Timothée Poisot[10], Hailey Robertson[1], Sadie J. Ryan[5], Stephanie N. Seifert[11], David Simons[12], Amanda Vicente-Santos[3], Chelsea L. Wood[13], Ellie Graeden[1,14], and Colin J. Carlson[1,2,15,*]

1. Center for Global Health Science and Security, Georgetown University
2. Department of Biology, Georgetown University
3. School of Biological Sciences, University of Oklahoma
4. Center for Vector-borne Infectious Diseases, Department of Microbiology, Immunology, and Pathology, Colorado State University
5. Emerging Pathogens Institute, University of Florida
6. Blue Tiger, LLC
7. Department of Ecological Dynamics, Leibniz Institute for Zoo and Wildlife Research
8. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
9. Wildlife Conservation Society, Health Program
10. Département de Sciences Biologiques, Université de Montréal
11. Paul G. Allen School for Global Health, University of Washington
12. Centre for Emerging, Endemic and Exotic Diseases, Royal Veterinary College
13. School of Aquatic and Fishery Sciences, University of Washington
14. Massive Data Institute, Georgetown University
15. Department of Epidemiology of Microbial Diseases, Yale University School of Public Health

* Correspondence should be directed to colin.carlson@georgetown.edu

Prepared as an Article for *Scientific Data*

Abstract

Thousands of scientists and practitioners conduct research on infectious diseases of wildlife. Rapid and comprehensive data sharing is vital to the transparency and actionability of their work, but unfortunately, most efforts designed to publically share these data are focused on pathogen determination and genetic sequence data. Other facets of existing surveillance data – particularly negative results – are often withheld or, at best, summarized in a descriptive table with limited metadata. As a result, very few datasets on wildlife disease dynamics over space and time are publicly available for synthesis research or applied uses in conservation or public health. Here, we propose a minimum data and metadata reporting standard for wildlife disease studies. Our checklist identifies a minimum set of 30 fields required to standardize and document a dataset consisting of records disaggregated to the finest possible spatial, temporal, and taxonomic scale. We illustrate how this standard is applied to an example study, which documented a novel alphacoronavirus found in bats in Belize. Finally, we outline best practices for how data should be formatted for optimal re-use, and how researchers can navigate potential safety concerns around data sharing.

Introduction

Infectious disease is a widely studied topic in wildlife biology and ecosystem science (1). Every year, countless scientific studies report new data on the prevalence of macroparasites (e.g., ticks and tapeworms) and microparasites (e.g., bacteria, viruses, and other classically defined "pathogens"), hereafter “parasites” for simplicity (2), in wild animals. These datasets can be used to test and reveal ecological principles, monitor the impacts of climate change and biodiversity loss, and even track emerging threats to human and ecosystem health.

Unfortunately, of the thousands of datasets produced every year, very few are shared publicly, and those that are often have limited potential for reuse. Many researchers still discard negative data, or focus almost exclusively on positive results: for example, a study might focus on the characterization of a single novel virus, while briefly acknowledging hundreds of samples that tested negative in a single sentence of the Methods or Results. Sometimes, raw data are at least summarized in a table that reports parasite prevalence for different combinations of host taxon, parasite taxon, and sampling location or period; these data often cannot be disaggregated back to the level of individual animals or test results. Many studies also fail to report essential metadata, such as primer sequences, sampling effort over space, time, or host taxa, or biologically-meaningful host characteristics such as body size (3,4). Collectively, these practices reduce the quality of the limited data that do become available for reuse and reanalysis, posing a significant challenge for synthesis research on infectious disease macroecology (5–7) or on the risks posed by emerging zoonotic and vector-borne diseases (8–10). In this regard, wildlife disease research has started to lag substantially behind other areas of the biological sciences, where open sharing of reusable primary data has become widely expected.

Building on a set of similar templates for sharing datasets related to arthropod disease vectors (11–13), we developed a minimum standard for wildlife infectious disease data that avoids unnecessary jargon, and that balances effort with detail and standardization with flexibility. Here, we document the standard’s development; show how the standard can be applied to a simple dataset; and suggest additional best practices for data sharing.

Methods

Our goal in this project was to develop guidelines for how researchers can share standardized and well-documented wildlife disease datasets, with a focus on capturing how sampling happened and what was found. We developed our data standard based on: (i)

experience conducting and publishing wildlife disease research, and collaborating with government programs doing the same; (ii) common practices already followed by most scientists in the literature when sharing disaggregated data, including the decisions made by major data sources such as the USAID PREDICT 2 project's data release (14); (iii) best practices for sharing ecological data that minimize room for error or loss of data (15–20); and (iv) interoperability with standards used by other platforms, such as the Global Biodiversity Informatics Facility (GBIF) (20). We also assumed that pathogen genetic sequence data and associated types (e.g., metatranscriptomes) are already widely archived on platforms like NCBI's GenBank and Sequence Read Archive (SRA), following a different set of best practices, and are unlikely to be stored in the same data structure as we describe here.

The guiding philosophy of the data standard is that researchers should share their raw data in a format that data scientists refer to as "rectangular data" or "tidy data" (21), where each row corresponds to a single measurement, here meaning the outcome of a diagnostic test. Tests, samples, and individual animals can each have many-to-many relationships due to common practices such as repeated sampling of the same animal, confirmatory tests or deeper sequencing of samples that test positive, and pooling of samples (sometimes from multiple animals and locations) for a single test. Based on this, there are three main categories of information collected: sample metadata, host animal metadata, and the parasite data itself, including both test results and any metadata characterizing a parasite once it has been detected (e.g., GenBank accession). The authors developed the fields associated with each of these categories through a process of iteration with real-world data, as part of the ongoing development of a new dedicated platform called the Pathogen Harmonized Observatory database (PHAROS: pharos.viralemergence.org). The PHAROS platform will be comprehensively documented in a separate manuscript.

Results

When to use the data standard

Before applying this standard, we encourage researchers to verify that their dataset describes wild animal samples that were examined for parasites, accompanied by metadata on the diagnostic methods used and the date and location of sampling. Some closely-related types of data are better stored in another format: for example, records of free-living macroparasites (e.g., tick dragging data) can be stored in Darwin Core format like any other biodiversity dataset (20,22), or can adhere to the MIReAD data standard for arthropod abundance data, which was designed with disease vector surveillance in mind (13). Similarly,

arthropod blood meal datasets can follow another recently-published data standard (12). Finally, environmental monitoring datasets (e.g., soil, water, or air microbiome metagenomics) not associated with a specific animal under direct or indirect observation should also be handled following other best practices (23,24).

The data standard

Our proposed data standard includes 30 core fields (nine related to sampling, 12 related to the host organism being sampled, and nine related to the parasite itself). The contents of these fields and their interpretation are described in Tables 1-3.

To illustrate how other scientists can use the minimum data standard, we present an example using a previously published dataset (25). The example dataset captures two records associated with a single vampire bat (BZ19-114) tested for coronaviruses in Belize in 2019: a rectal swab tested negative, while an oral swab tested positive, leading to the identification of a novel alphacoronavirus (Table 4). All mandatory and relevant fields are shown, and cells are only left blank if they do not apply (i.e., parasite identity and GenBank accession are always empty for negative test results). The data in Table 4 are only a subset of the full dataset, which is shared in full on the PHAROS platform (project: prjRPayEvMecN).

Removing and adding fields

Some datasets will not be able to meet a comprehensive standard for documentation. Wherever possible, we encourage researchers to leave fields blank, rather than remove them. For example, in some projects, limited funding or study protocols may preclude all captured animals from being sampled, or all samples from being tested. Researchers might therefore include a mix of records of animals or samples with no attached test data (i.e., leaving "Detection outcome" blank). Similarly, archival samples that are rescued from old projects, or older museum specimens that are sampled for parasites (26), may not always have complete date information, leading "Collection day" and "Collection month" to be left blank. We encourage researchers to adapt our data standard to their purposes, and as appropriate, to consider sharing their data in multiple applicable formats. For example, in the previous example, researchers might share their test results on the PHAROS platform, but share a more comprehensive record of all sampling in the study's supplemental materials.

Some datasets will also need to include additional fields capturing other kinds of information. For example, researchers might use an all-purpose "Notes" column to flag unusual records or non-standardized information about sampling (e.g., the circumstances under which a dead

animal was found, such as opportunistic roadkill collection). Similarly, in cases where findings are particularly sensitive for public health or economic reasons, researchers might even consider including some guidance for how to interpret them in the data itself: for example, the data shared by the USAID PREDICT 2 project includes a field called "Interpretation," which provides guidance such as this disclaimer on a positive test result: "[The virus detected in this sample] is the known ebolavirus, Bombali virus, detected in an Angolan free-tailed bat. This virus has previously been found in bats in Sierra Leone as part of the PREDICT project. Further characterization is ongoing to understand the zoonotic potential of this virus."

Best practices for sharing (and withholding) data

When using the data standard, we suggest that researchers should follow scientific conventions and best practices for data science, such as: reporting measurements in metric units; reporting taxonomic information at the most granular level possible for both the host and parasite; and leaving empty and non-applicable cells blank, rather than assigning a placeholder such as "NA" (27). Researchers should also ensure that their manuscript comprehensively describes all important aspects of sampling methodology, such as the circumstances (e.g., systematic and planned sampling versus opportunistic collection of unusual carcasses), how animals were identified (e.g., expert opinion versus barcoding), and how samples were prepared (e.g., specific products or kits used, or specific details about the methods used in parasitological dissections). None of these are likely to differ for each individual row of data, and so we exclude these from the template, but interpreting a study's data correctly may still depend on these data being available. Researchers should also make sure that their study documents any relevant epidemiological observations (e.g., unusual disease presentation, or sewage discharge or farms nearby). Finally, whenever possible, researchers should also share all sequence data in an open repository.

As with other kinds of biodiversity data (28,29), sharing high-resolution wildlife disease data can sometimes be unsafe or inadvisable. For example, sharing the location of a bat roost where viruses have been detected may lead to culling, which in turn puts local communities at greater risk of exposure (30,31). There may also be biosafety or biosecurity risks associated with location data, depending on the characteristics of the parasite in question: for example, anthrax spores can persist at a carcass site for several years (32,33). In sensitive cases, researchers could consider truncating longitude and latitude values, or potentially, jittering records with random noise, and should then carefully and clearly document the obfuscation process; guidance on this practice exists for other kinds of biodiversity data (34). In some cases, this may still be insufficient to prevent malicious use (35). In high-risk

cases, journal editors should work closely with authors to ensure that neither the manuscript itself nor any supplementary data have a significant potential to cause harm.

Discussion

Here, we propose a data standard for wildlife infectious disease studies. With minimal modifications, the same template could also be used for related types of data, such as records of plant diseases, or infections in captive animal populations such as zoos and wildlife sanctuaries. However, other types of spatiotemporal disease data may already have associated best practices and dedicated or otherwise well-suited repositories. For example, disaggregated but carefully de-identified human infectious disease data can be shared in epidemic settings on the Global.health platform (36); host, vector, and parasite occurrence data can also all be documented in Darwin Core format and shared in GBIF (37–39).

We encourage researchers to consider adopting this minimum standard when publishing research that uses wildlife disease data. To encourage this practice, blank templates (in both .xlsx and .csv format) are available both as supplementary files to this manuscript and on a public GitHub repository (github.com/viralemergence/pharos-standard). We suggest that researchers should share their formatted data as a supplemental file accompanying a publication, or better yet, deposit their data in a repository such as Figshare, Dryad, or Zenodo. A modified version of this data standard is also implemented in the PHAROS platform, which allows researchers to manage and publish their data on a platform built specifically for wildlife disease research and surveillance. Sharing datasets on this dedicated platform makes them more findable than on all-purpose repositories, while still providing a system for data citations based on dataset- and download-specific identifiers. Researchers are also encouraged to share data on PHAROS before or independent of publication, especially in cases where negative data might not be publishable, or where timely sharing of findings might be particularly relevant to public health or conservation.

Whether or not researchers share their data on the PHAROS platform, we hope they will consider using this minimum data standard to ensure their data are findable, accessible, interoperable, and reusable (FAIR) by other scientists (40). Doing so will also help studies meet the minimum requirements for data sharing now adopted by most journals and scientific funders (from which a surprising number of studies still actively seek exception). Progress toward open science will make wildlife disease research a richer and more rigorous field, leading to better insights about emerging threats to human and animal health.

Data Availability

The example dataset and blank templates are available from Github.

Code Availability

No code is used in this manuscript.

Acknowledgements

This work was supported by an NSF Biology Integration Institute grant (NSF DBI 2021909 and 2213854). We also thank countless colleagues for conversations and work that shaped this data standard, including Noam Ross and the team at the USAID Development Data Library.

Figures and Tables

Table 1. Sampling metadata.

Variable	Descriptor
Sample ID	A researcher-generated unique ID for the sample: usually a unique string of both characters and integers (e.g., "OS BZ19-114" to indicate an oral swab taken from animal BZ19-114; see worked example below), to avoid conflicts that can arise when datasets are merged with number-only notation for samples. Ideally, sample names should be kept consistent across all online databases and physical resources (e.g., museum collections or project-specific sample archives).
Animal ID	A researcher-generated unique ID for the individual animal from which the sample was collected: usually a unique string of both characters and integers (e.g., "BZ19-114" to indicate animal 114 sampled in 2019 in Belize). Ideally, animal names should again be kept consistent across online databases and physical resources.
Latitude	Latitude of the collection site in decimal format.
Longitude	Longitude of the collection site in decimal format.
Spatial uncertainty *(optional)*	Coordinate uncertainty from GPS recordings, post-hoc digitization, or systematic alterations (e.g., jittering or rounding) expressed in meters.
Collection day	The day of the month on which the specimen was collected.
Collection month	The numeric month in which the specimen was collected.
Collection year	The year in which the specimen was collected.
Collection method and/or tissue	The technique used to extract the sample and/or the tissue from which the sample was extracted (e.g., "oropharyngeal swab")

Table 2. Host identification and traits.

Variable	Descriptor
Host identification	The Linnaean classification of the animal from which the sample was collected, reported at the lowest possible level (ideally, species binomial name: e.g., "Odocoileus virginianus" or "Ixodes scapularis"). As necessary, researchers may also include an additional field indicating when uncertainty exists in the identification of the host organism (see "Adding new fields").
Organism sex *(optional)*	The sex of the individual animal from which the sample was collected.
Dead or alive *(optional)*	The state of the individual animal from which the sample was collected, at the time of sample collection.
Health notes *(optional)*	Any additional (unstructured) notes about the state of the animal, such as disease presentation.
Life stage *(optional)*	The life stage of the animal from which the sample was collected (as appropriate for the organism) (e.g., "juvenile", "adult").
Age *(optional)*	The numeric age of the animal from which the sample was collected, at the time of sample collection, if known (e.g., in monitored populations).
Age units *(optional)*	The units in which age is measured (usually years).
Mass *(optional)*	The mass of the animal from which the sample was collected, at the time of sample collection.
Mass units *(optional)*	The units that mass is recorded in (e.g., "kg").
Length *(optional)*	The numeric length of the animal from which the sample was collected, at the time of sample collection.
Length measurement *(optional)*	The axis of measurement for the organism being measured (e.g., "snout-vent length" or just "SVL"; "wing length"; "primary feather").
Length units *(optional)*	The units that length is recorded in (e.g., "mm").

Table 3. Detection methods and parasite identification.

Variable	Descriptor
Detection target	The taxonomic identity of the parasite being screened for in the sample. This will often be coarser than the identity of a specific parasite identified in the sample: for example, in a study screening for novel bat coronaviruses, the entire family *Coronaviridae* might be the target; in a parasite dissection, the targets might be Acanthocephala, Cestoda, Nematoda, and Trematoda. For deep sequencing approaches (e.g., metagenomic and metatranscriptomic viral discovery), researchers should report each alignment target used as a new "test" to maximize reporting of negative data, or alternatively, select a subset that reflect specific study objectives and the focus of analysis (e.g., specific viral families).
Detection method	The type of test performed to detect the parasite or parasite-specific antibody (e.g., 'qPCR', 'ELISA').
Primer sequence *(optional)*	The sequence of both forward and reverse primers used to identify the sample (e.g., "forward 5' CDCAYGARTTYTGYTCNCARC 3' ; reverse 5' RHGGRTANGCRTCWATDGC 3'") or just the name of a commonly used gene target (particularly if citation information is given).
Primer citation *(optional)*	Citation for the primer being used.
Detection outcome	The test result (i.e., "positive", "negative", or "inconclusive"). To avoid ambiguity, these specific values are suggested over numeric values ("0" or "1").
Detection measurement *(optional)*	Any numeric measurement of parasite detection that is more detailed than simple positive or negative results (e.g., viral titer, parasite counts, sequence reads).
Detection measurement units *(optional)*	Units for quantitative measurements of parasite intensity or test results (e.g., "Ct", "TCID50/mL", or "parasite count").
Parasite identification	The identity of a parasite detected by the test, if any, reported to the lowest possible taxonomic level, either as a Linnaean binomial classification or within the convention of a relevant taxonomic authority (e.g., "Borrelia burgdorferi" or "Zika virus"). Parasite identification may be more specific than detection target.
GenBank accession *(optional)*	The GenBank accession for any parasite genetic sequence(s), if appropriate. Researchers may also add additional / other fields as appropriate, such as for other genomic sequence data platforms (e.g., GISAID).

Table 4. An example of wildlife disease records following the minimum data standard.

Sample ID	Animal ID	Latitude	Longitude	Collection day	Collection month	Collection year	Collection method
OS BZ19-95	BZ19-114	17.76425974	-88.65209879	23	04	2019	Oral swab
RS BZ19-95	BZ19-114	17.76425974	-88.65209879	23	04	2019	Rectal swab

Host identification	Organism sex	Dead or alive	Life stage	Mass	Mass units
Desmodus rotundus	male	alive	subadult	0.023	kg
Desmodus rotundus	male	alive	subadult	0.023	kg

Detection target	Detection method	Primer sequence	Primer citation	Detection outcome	Parasite identification	GenBank accession
Coronaviridae	semi-nested PCR	RdRp	doi:10.3390/v9120364	positive	Alphacoronavirus	OM240578
Coronaviridae	semi-nested PCR	RdRp	doi: 10.3390/v9120364	negative		

References

1. McCallen, E. *et al.* Trends in ecology: shifts in ecological research themes over the past four decades. *Front. Ecol. Environ.* **17,** 109–116 (2019).
2. Lafferty, K. D. & Kuris, A. M. Trophic strategies, animal diversity and body size. *Trends Ecol. Evol.* **17,** 507–513 (2002).
3. Albery, G. F., Sweeny, A. R., Becker, D. J. & Bansal, S. Fine-scale spatial patterns of wildlife disease are common and understudied. *Funct. Ecol.* **36,** 214–225 (2022).
4. Wood, C. L. & Lafferty, K. D. How have fisheries affected parasite communities? *Parasitology* **142,** 134–144 (2015).
5. Stephens, P. R. *et al.* The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. *Ecol. Lett.* **19,** 1159–1171 (2016).
6. Halpern, B. S. *et al.* Priorities for synthesis research in ecology and environmental science. *Ecosphere* **14,** (2023).
7. Cohen, J. M., Sauer, E. L., Santiago, O., Spencer, S. & Rohr, J. R. Divergent impacts of warming weather on wildlife disease risk across climates. *Science* **370,** (2020).
8. Cohen, L. E., Fagre, A. C., Chen, B., Carlson, C. J. & Becker, D. J. Coronavirus sampling and surveillance in bats from 1996–2019: a systematic review and meta-analysis. *Nature Microbiology* **8,** 1176–1186 (2023).
9. Becker, D. J., Crowley, D. E., Washburne, A. D. & Plowright, R. K. Temporal and spatial limitations in global surveillance for bat filoviruses and henipaviruses. *Biol. Lett.* **15,** 20190423 (2019).
10. Tolsá, M. J., García-Peña, G. E., Rico-Chávez, O., Roche, B. & Suzán, G. Macroecology of birds potentially susceptible to West Nile virus. *Proc. Biol. Sci.* **285,** 20182178 (2018).
11. Wu, V. Y. *et al.* A minimum data standard for vector competence experiments. *Sci Data* **9,** 634 (2022).
12. Wehmeyer, M. L., Sauer, F. G. & Lühken, R. A minimum data standard for reporting host-feeding patterns of vectors. *ResearchSquare* (2024). doi:10.21203/rs.3.rs-3896902/v1.
13. Rund, S. S. C. *et al.* MIReAD, a minimum information standard for reporting arthropod abundance data. *Sci Data* **6,** 40 (2019).
14. PREDICT Consortium. PREDICT Emerging Pandemic Threats Project. Dataset. USAID Development Data Library (2021). https://data.usaid.gov/d/tqea-hwmr.
15. Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D. & Peres-Neto, P. Ecological Data Should Not Be So Hard to Find and Reuse. *Trends Ecol. Evol.* **34,** 494–496 (2019).
16. Guralnick, R., Walls, R. & Jetz, W. Humboldt Core - toward a standardized capture of

biological inventories for biodiversity monitoring, modeling and assessment. *Ecography* **41,** 713–725 (2018).
17. Augustine, S. P., Bailey-Marren, I., Charton, K. T., Kiel, N. G. & Peyton, M. S. Improper data practices erode the quality of global ecological databases and impede the progress of ecological research. *Glob. Chang. Biol.* **30,** e17116 (2024).
18. Costello, M. J. & Wieczorek, J. Best practice for biodiversity data management and publication. *Biol. Conserv.* **173,** 68–73 (2014).
19. Keller, A. *et al.* Ten (mostly) simple rules to future-proof trait data in ecological and evolutionary sciences. *Methods Ecol. Evol.* **14,** 444–458 (2023).
20. Wieczorek, J. *et al.* Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* **7,** e29715 (2012).
21. Wickham, H., Çetinkaya-Rundel, M. & Grolemund, G. *R for Data Science*. ('O'Reilly Media, Inc.', 2023).
22. Paull, S. H., Thibault, K. M. & Benson, A. L. Tick abundance, diversity and pathogen data collected by the National Ecological Observatory Network. *GigaByte* **2022,** gigabyte56 (2022).
23. Vangay Pajau *et al.* Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities. *mSystems* **6,** 10.1128/msystems.01194–20 (2021).
24. Huttenhower, C., Finn, R. D. & McHardy, A. C. Challenges and opportunities in sharing microbiome data and analyses. *Nat Microbiol* **8,** 1960–1970 (2023).
25. Becker, D. J. *et al.* Serum proteomics identifies immune pathways and candidate biomarkers of coronavirus infection in wild vampire bats. *bioRxiv* 2022.01.26.477790 (2022). doi:10.1101/2022.01.26.477790.
26. Wood, C. L. *et al.* A reconstruction of parasite burden reveals one century of climate-associated parasite decline. *Proc. Natl. Acad. Sci. U. S. A.* **120,** e2211903120 (2023).
27. White, E. *et al.* Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution* **6,** 1–10 (2013).
28. Tulloch, A. I. T. *et al.* A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nat Ecol Evol* **2,** 1209–1217 (2018).
29. Lunghi, E., Corti, C., Manenti, R. & Ficetola, G. F. Consider species specialism when publishing datasets. *Nat Ecol Evol* **3,** 319 (2019).
30. Shapiro, J. T. *et al.* Setting the Terms for Zoonotic Diseases: Effective Communication for Research, Conservation, and Public Policy. *Viruses* **13,** (2021).
31. Amman, B. R. *et al.* Marburgvirus resurgence in Kitaka Mine bat population after extermination attempts, Uganda. *Emerg. Infect. Dis.* **20,** 1761–1764 (2014).

32. Carlson, C. J. *et al.* Spores and soil from six sides: interdisciplinarity and the environmental biology of anthrax (Bacillus anthracis). *Biol. Rev. Camb. Philos. Soc.* **93,** 1813–1831 (2018).
33. Barandongo, Z. R. *et al.* The persistence of time: the lifespan of Bacillus anthracis spores in environmental reservoirs. *Res. Microbiol.* **174,** 104029 (2023).
34. Chapman, A. D. & Grafton, O. Guide to best practices for generalising sensitive/primary species occurrence-data. Version 1.0. (2008). at <https://repository.oceanbestpractices.org/handle/11329/605>
35. Beery, S. & Bondi, E. Can poachers find animals from public camera trap images? *arXiv* doi:10.48550/arXiv.2106.11236.
36. Benjamin, A. et al. Global.health: a scalable plaform for pandemic data integration, analytics, and preparedness. *ResearchSquare* (2022). doi:10.21203/rs.3.rs-1528783/v1.
37. Salim, J. A., Seltmann, K., Poelen, J. & Saraiva, A. Indexing Biotic Interactions in GBIF data. *Biodivers. Inf. Sci. Stand.* **6,** (2022).
38. Astorga, F. *et al.* Biodiversity data supports research on human infectious diseases: Global trends, challenges, and opportunities. *One Health* **16,** 100484 (2023).
39. Edmunds, S. C. *et al.* Publishing data to support the fight against human vector-borne diseases. *Gigascience* **11,** (2022).
40. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016).