SCIENTIFIC DATA

OPEN

ARTICLE

MIReAD, a minimum information standard for reporting arthropod abundance data

Received: 19 October 2018
Accepted: 20 March 2019
Published online: 25 April 2019

Samuel S. C. Rund[1], Kyle Braak[3], Lauren Cator[4], Kyle Copas[3], Scott J. Emrich[5], Gloria I. Giraldo-Calderón[1,2], Michael A. Johansson[6,7], Naveed Heydari[8], Donald Hobern[3], Sarah A. Kelly[9], Daniel Lawson[9], Cynthia Lord[10], Robert M. MacCallum[9], Dominique G. Roche[11], Sadie J. Ryan[12,13,14], Dmitry Schigel[3], Kurt Vandegrift[15], Matthew Watts[4], Jennifer M. Zaspel[16] & Samraat Pawar[4]

Arthropods play a dominant role in natural and human-modified terrestrial ecosystem dynamics. Spatially-explicit arthropod population time-series data are crucial for statistical or mathematical models of these dynamics and assessment of their veterinary, medical, agricultural, and ecological impacts. Such data have been collected world-wide for over a century, but remain scattered and largely inaccessible. In particular, with the ever-present and growing threat of arthropod pests and vectors of infectious diseases, there are numerous historical and ongoing surveillance efforts, but the data are not reported in consistent formats and typically lack sufficient metadata to make reuse and re-analysis possible. Here, we present the first-ever minimum information standard for arthropod abundance, Minimum Information for Reusable Arthropod Abundance Data (MIReAD). Developed with broad stakeholder collaboration, it balances sufficiency for reuse with the practicality of preparing the data for submission. It is designed to optimize data (re)usability from the "FAIR," (Findable, Accessible, Interoperable, and Reusable) principles of public data archiving (PDA). This standard will facilitate data unification across research initiatives and communities dedicated to surveillance for detection and control of vector-borne diseases and pests.

Introduction

Arthropods play a dominant role in the dynamics of practically all natural and human-modified terrestrial ecosystems[1–3] and have significant economic and health effects. For example, certain insects provide significant economic benefits (*e.g.* pollination) exceeding $57 billion a year to the United States alone[4]. Invasive insects, however, cost an estimated $70 billion dollars per year globally[5] and insect pests may reduce agricultural harvests by up to 16%, with an equal amount of further losses of harvested goods[6]. Particularly noteworthy is a subset of arthropods

[1]VectorBase, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA. [2]Universidad Icesi, Facultad de Ciencias Naturales, Calle 18 No. 122-135, Cali, Colombia. [3]Global Biodiversity Information Facility (GBIF) Secretariat, Copenhagen, Denmark. [4]Department of Life Sciences, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, Berkshire, United Kingdom. [5]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA. [6]Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, 1324 Calle Cañada, San Juan, PR, USA. [7]Department of Epidemiology, Harvard School of Public Health, 677 Huntington Ave, Boston, MA, USA. [8]Center for Global Health and Translational Science, State University of New York Upstate Medical University, Syracuse, NY, USA. [9]VectorBase and Vector Immunogenomics and Infection Laboratory, Department of Life Sciences, Imperial College London, London, United Kingdom. [10]Florida Medical Entomology Lab, University of Florida-IFAS, Vero Beach, FL, USA. [11]Institute of Biology, University of Neuchâtel, 2000, Neuchâtel, Switzerland. [12]Quantitative Disease Ecology and Conservation Lab, Department of Geography, University of Florida, Gainesville, FL, USA. [13]Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA. [14]College of Life Sciences, University of Kwa-Zulu Natal, Durban, South Africa. [15]Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA, USA. [16]Department of Zoology, Milwaukee Public Museum, Milwaukee, WI, USA. Correspondence and requests for materials should be addressed to S.S.C.R. (email: srund@nd.edu)

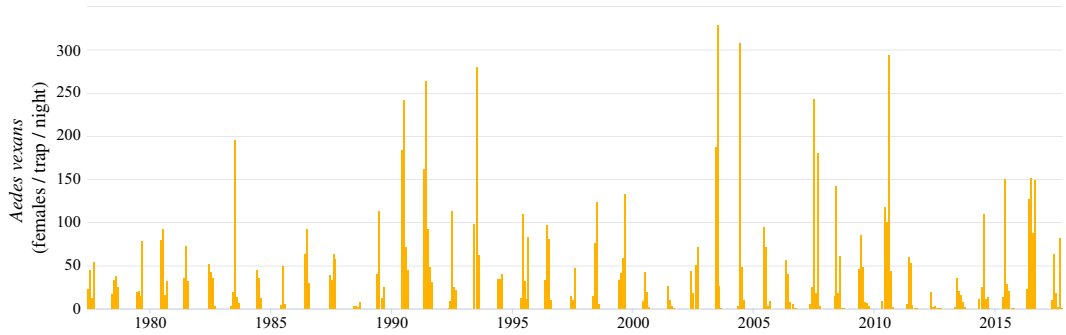


Fig. 1 Population abundance time-series example. From New Jersey light trap mosquito surveillance performed by the Iowa State University Medical Entomology Laboratory from 1977-2017[61,62]. Data available for download on VectorBase[50].

that are disease vectors, transmitting pathogens to and between animals as well as plants. Vector borne diseases cause billions of dollars in crop and livestock losses, every year[7–9]. In humans, vector borne diseases account for more than 17% of all infectious diseases (*e.g.* malaria, Chagas, dengue, and leishmaniasis, Zika, West Nile, Lyme disease, and sleeping sickness), with hundreds of thousands of deaths, hundreds of millions of cases, and billions of people at risk, annually[10,11].

The current economic and health burden of arthropod pests, exacerbated by invasive species, and uncertain effects of climate change[12,13], has driven significant research programs and data collection efforts. These include crop pest, mosquito, and tick survey and reporting initiatives[14–18], citizen science projects[19–21], and digitization of museum specimen data[22,23], all yielding a rich and growing trove of field-based data spanning multiple spatial and temporal scales. The monitoring of arthropod abundance (*e.g.* Fig. 1) in different disciplines (*e.g.*, biodiversity research, pest-control assessment, vector borne disease monitoring, and pollination research) has similar objectives — to quantify abundance, phenology and geographical ranges of target arthropod species — and entails similar techniques. However, the data produced by these various efforts are often not reusable, or comparable to similar data, as they are typically not recorded in a standard format (*e.g.* Darwin Core), or do not provide adequate metadata. In contrast, the advent of journal-mandated deposition of data from high-throughput technologies (*e.g.* NCBI and GenBank), data and code sharing, and other practices to improve transparency and reusability of research results are increasing rapidly across the sciences[24–29]. Furthering these advances through standardization and public archiving of arthropod abundance data can bring significant benefits, including (1) supporting empirical parameterization and validation of mathematical models (*e.g.* of pest or disease emergence and spread), (2) validation of model predictions, (3) reduction in the duplication of expensive empirical research, and (4) revealing new patterns and questions through meta-analyses[11,30–33]. This will also lead to substantial public benefit through improved human, animal, plant, and ecosystem health, and reduced economic costs.

One of the key impediments to the re-use of these data is the lack of adequate metadata or data descriptors (*i.e.* data about the data)[34–37]. In general, for data to be most valuable to the scientific community, they should meet the FAIR Principles – they should be Findable, Accessible, Interoperable and Reusable – and delineate the key components of good data management and stewardship practices[38,39]. Data are Findable and Accessible when they are archived and freely downloadable from an online public data repository that is indexed and easily searchable. Interoperability and reusability describe the ease with which humans or computer programs can understand the data (*e.g.* via metadata) and explore/re-use them across a variety of non-proprietary platforms. Even when data are available, metadata for arthropod abundance data are often absent or not readily interpretable, limiting their reusability at a fundamental level.

Results

A minimum information standard for arthropod abundance data. Here, we present a Minimum Information for Reusable Arthropod Abundance Data (MIReAD) standard for reporting primarily longitudinal (repeated, temporally explicit) field-based collections of arthropods. 'MIReAD' also evokes 'Myriad,' a countless or extremely great number. Abundance is measured and reported in different ways, and MIReAD fields have been designed to allow researchers to capture this complexity. Examples 1–4 (which can be found in Figshare[40]), provide examples for how to report such different types of abundances. However, we do not encourage the reporting of relative population abundances since these are not raw data as such, but derived values. One might argue that this could lead to (and it probably will) loss of information, but we argue that the reporting of raw abundance or occurrence data is non-negotiable if these data are to be reused. For example, incorrect statistical methods to aggregate data, such as taking the arithmetic mean for skewed abundance data across samples or replicates are not uncommon, and we wish to discourage such practices. In the same manner as has been developed in other biological disciplines[41–46], this standard is "minimum" because it defines the necessary minimal information required to understand and reuse a dataset without consulting any further persons, text, materials, or methods[47]. MIReAD is designed to facilitate data archiving efforts of publishers and field researchers. It is not a data model [the explicit definition of data field names, data formats (*e.g.*, for dates and GPS locations)] and therefore does not define controlled vocabularies, or specific field titles, but should be easy to understand and interpret by the wider scientific community[47].

MIReAD is separated into two components, metadata and data. For each component, we provide a description of the information that should be included, recommendations for how to make that information as useful as possible, and examples. The metadata component (Online-only Table 1) includes information for the origin of the data set (*e.g.* study information and licensing for usage). The second component (Online-only Table 2) lists and describes specific data fields that should be included in data collection sheets. We also provide recommendations and examples to demonstrate how these recommendations can be implemented. MIReAD was designed to match the data that are generally collected by academic researchers and surveillance initiatives, and can serve as a checklist for important information that needs to be recorded but is often unintentionally omitted (*e.g.* Fig. 2a). By adhering to MIReAD standards, omissions and ambiguity can be avoided even if the data are shared in different formats (Fig. 2b,c). Finally, we identify common problems likely to be encountered across all the MIReAD metadata and data fields, and data quality standards that can be employed to avoid confusion (Box 1).

Box 1 Data quality standards.

No abbreviations.

Abbreviations (including in columns names) are ambiguous, with the exception of measurement units (e.g. centigrade and meters).

No external legend/key files.

While repetitive, all data should be explicitly given within the data table. Separate files mapping ID numbers to GPS locations, full species names, *etc*. should be avoided. In addition, rich metadata are essential for good data discovery and reuse.

Unambiguous dates.

Because of country-level differences in date formats, data should be reported unambiguously with four-digit years, and months provided alphabetically and not numerically (e.g. 4-Jun-2017 or Nov 12, 2015,) or by using ISO 8601 date format (YYYY-MM-DD, e.g. 2019-01-27)

Machine-readable file formats.

Data should be provided in non-proprietary machine-readable formats such as comma-separated text files. PDFs and multiple spreadsheets in the same document should be avoided.

No font styling or subsection headings.

Formatting (color, bold, italics, subscripts, sheet tab names, etc.) should not be required for understanding the data. Subsection headings should not be required to understand data; every line of data should be interpretable in isolation from any other line of data.

Highest precision possible.

Data should be provided at the highest temporal, spatial, numerical, and taxonomic resolution available. If location (e.g., geographical coordinate) data need to be presented at a lower resolution than available for privacy reasons, this should be made clear in the submission in Study Information (Resource Metadata; Online-only Table 1).

Language.

Once data are ready to be deposited/submitted, all fields and data should be preferably written in English. This will allow researchers and data curators worldwide to understand and reuse the data. Use of other languages is better than not publishing data. Please avoid introducing data reuse barriers through incomplete translation. For example, avoid non-English field names in an English-language submission.

Examples. Below we provide three examples to illustrate MIReAD-compliant data (linked to examples 1–4 in Figshare[40], respectively). Researchers can use these data sheets as a basis for formatting their own data. In these examples, note that all data meet the data quality standards of Box 1; are adequately described, have columns labeled, *etc*. to eliminate ambiguity (even if the data appear repetitive; for example, the sex and life stage are repeated in every row). Examples 1 and 2 should be sufficient for most data generators. Examples 3–4 demonstrate more complex data collection scenarios.

Long-format trapping data. Each row captures count data for a single species' occurrence in a given sampling event. This illustrates an example of the most common mosquito collection protocol (MIReAD_example_1.csv[40]). Also see Fig. 2b.

Wide format trapping data. Each row captures count data from a given sampling event. Each identified taxonomic group is identified in a separate column. An 'additional sample information' field, 'sub-location,' has been added to describe the various locations around the village where collections were made (MIReAD_example_2.csv[40]). This illustrates an example of adult mosquito populations that have been tracked over time and in specific locations. Also see Fig. 2c.

Complex trapping data scenarios. Tick surveillance performed using tick drags and flags and collections of ectoparasites on trapped mice. The tick drags/flags report three life stages independently (adult, larvae, and nymph) (MIReAD_example_3.csv[40]). Larvae are only identified to the genus, while adults and nymphs are identified to the species. Observations of different life stages and sexes are preferably documented in separate records. A Sample Name is used to help link these records (but would not be necessary.) The mouse survey uses an additional sample information field to record the sex of the trapped mouse from which the parasites were collected (MIReAD_example_4.csv[40]).

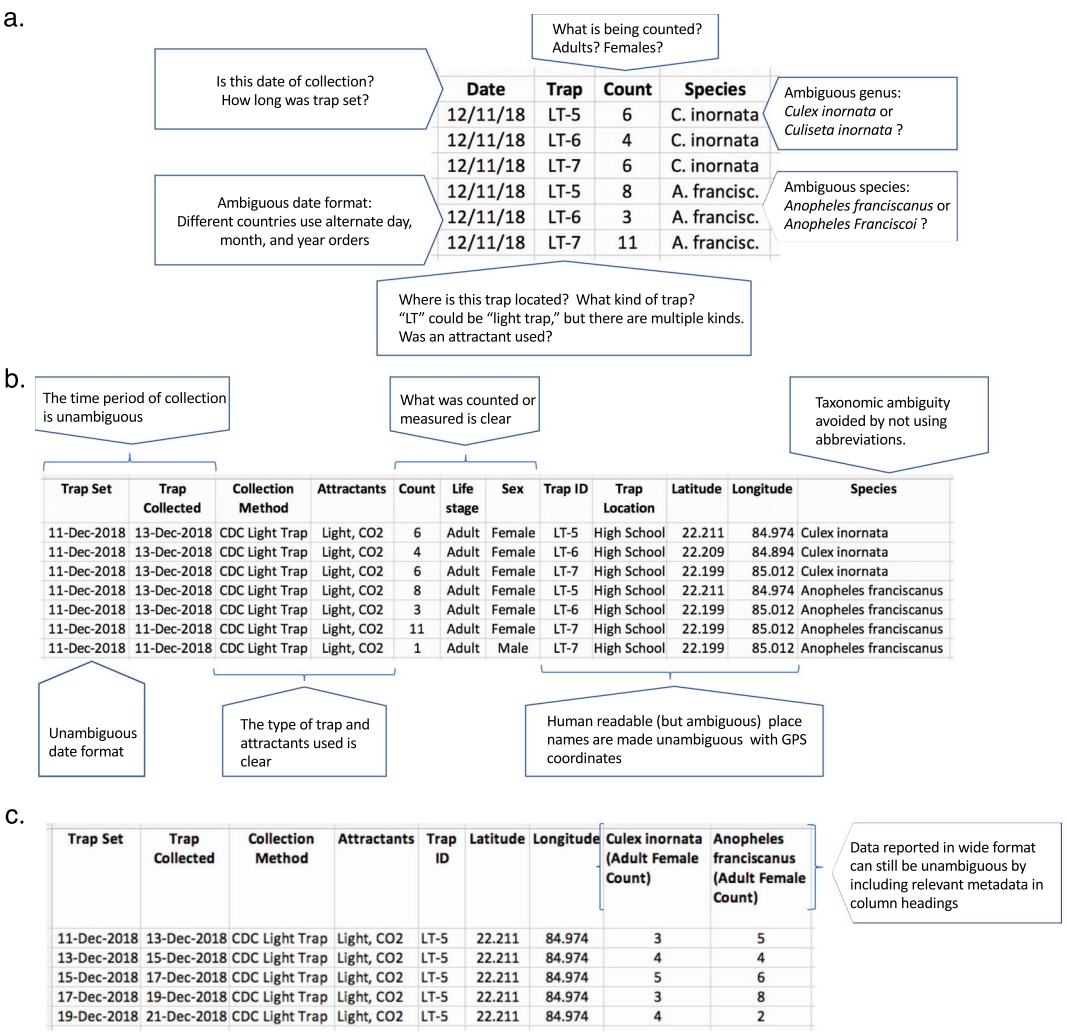


Fig. 2 MIReAD reduces data ambiguity. (**a**) Seemingly clean data can still lack key information or have ambiguous metadata, hindering data reuse. (**b**) MIReAD compliant data includes the metadata necessary for data reuse and removes ambiguity. (**c**) Note data can be formatted differently, but still be MIReAD compliant such as by presenting data in a wide format.

Discussion

MIReAD as the path to FAIR data principles. We designed MIReAD to achieve a balance between standards that are too onerous for data generators with guidelines that are sufficient to ensure at least minimal reusability[31,41]. It balances a perfectly formatted and reusable dataset with all necessary metadata in a consistent format (but comes with a high burden on the dataset generator) versus a dataset that is unusable or re-usable due to missing or incomplete metadata. MIReAD allows for a relatively easy standardization, as it ensures all necessary collection metadata is present in an unambiguous manner. By not mandating any particular field name, field order, or controlled vocabulary terms, we will in fact gain traction from other more rigorous (and thus more onerous) data models, for which lack of minimum standards in data are often a first and major hurdle. In striking this balance, we note that MIReAD focuses on capturing information on 'what' was done, rather than 'why'. We acknowledge that for some use cases, this may hinder reusability[48] but for the majority of cases, where the results of the original data can be interpreted without understanding the rationale, providing data in MIReAD format will be sufficient for data reuse.

Like all minimum standards, MIReAD only aims at ensuring data 'Reusability'. However, ultimately this will promote the implementation of data models, and controlled vocabularies (*e.g.*, the Darwin Core[49]). Data models enable 'Interoperability', and in turn facilitate structured databases, public repositories, and development of data analysis tools[47,50]. Deposition in open databases make data 'Findable' and 'Accessible'[51–53]. MIReAD compliant data contain sufficient information for established aggregators/databases such as VectorBase and SCAN (Symbiota Collections of Arthropods Network[54]) to process and store the data in a standardized data model [*e.g.*, Darwin Core, a widely used universal data standard that supports opportunistic observation and collection data (occurrence core) as well as presence/absence and abundance data collected using strict and documented methodology (event core)[49]], and ultimately facilitate data transfer to even more comprehensive biodiversity databases. For example, GBIF contains over one billion species occurrence records, from thousands of environmental,

MIReAD Field	Corresponding GBIF metadata fields
Contact details	contact
General description of the experiment/collection set	designDescription; sampling
Citations	citation
Species Identification Method	designDescription
Not present vs zero information	samplingDescription
GPS obfuscation information	geographicDescription geodeticDatum
Data usage information	intellectualRights

Table 1. The MIReAD data fields and their mapping to the GBIF metadata profile. See GBIF[63] for more information.

MIReAD Field	Corresponding DarwinCore fields
Start Time (for collection)	eventTime
End Time (for collection)	eventTime
Location	A number of fields under Location See: http://rs.tdwg.org/dwc/terms/#location
Collection method	samplingProtocol
Collection attractants	samplingProtocol
Collection area	samplingEffort
Taxonomy	A number of fields under Taxon See http://rs.tdwg.org/dwc/terms/#taxon
Unit(s) of measurement and observation	sampleSizeUnite
Value	sampleSizeValue
Additional sample information	fieldNotes
	eventRemarks
SampleID	eventID
Sample Name	*e.g.* fieldNumber for individual observations; see also SampleID above for the field names for the complex samples

Table 2. The MIReAD data fields and their mapping to the Darwin Core data standard. See Wieczorek *et al.*[64] for more information.

ecological, and natural resource investigations, including research on Arthropoda in numerous ecological and monitoring projects, allowing for study of changes and trends in populations[53]. Indeed, in Tables 1, 2, we provide an example of the mapping of data fields from MIReAD to DarwinCore and GBIF. In this way, MIReAD opens the door to FAIR data and more sophisticated methods to integrate data across many scales.

Benefits to field researchers. It is essential that the benefits of a minimal data standard extend not just to data re-users, but also to the researchers who collect and generate data in the first place. MIReAD provides a framework for data preparation that can help scientists achieve recognized professional merit for sharing data such as increased citation rates, academic recognition, opportunities for co-authorship, and new collaborations [sensu Roche *et al.*[31]]. Large, deposited data sets can now themselves be standalone, citable "data papers" (*e.g.*[55–57]) or even depositions without any traditional manuscript (but as an authored 'digital product,' with persistent identifiers, such as a DOI number), if desired. Data sets are increasingly recognized as valuable research outputs that count towards academic recognition and professional advancement (*e.g.* grants, interviews, and tenure). For example, several funders (*e.g.* United States National Science Foundation and Swiss National Science Foundation) have adopted or are in the process of adopting the Declaration on Research Assessments (DORA)[58], offering further opportunities for data generators to gain recognition and publication credit for their work[59]. Also, an increasing number of funders are mandating public data access, and detailed data management plans are often required even at the grant proposal stage. Therefore, reporting data according to MIReAD will provide a basis for stipulating archival formats. We also note that by storing data in MIReAD format, data generators can assure that their data contains all the necessary metadata for their own internal use. As time passes, research staff, sampling protocols, and sampling locations change, and thus the recording of minimal information ensures long-term reusability of data.

Furthermore, many data generators are also data users. Developing analyses that rely on standardized fields can facilitate the development of generalized analytical tools that can be easily extended to datasets beyond those that were collected by a single individual or lab. In this way, they can enable extensions of work that would otherwise not happen, such as comparisons of population dynamics in different locations or assessments of interspecies interactions. Adopting MIReAD can, therefore, both help data generators reap the benefits of sharing data they have collected and enable them to more readily leverage data collected by others.

Further MIReAD applications and extensions. The creation of minimum information standards for these types of databases facilitates analyses of data at scales that cannot be attained by a single individual or lab group. Linking records to additional information also extends the utility of these data to address population level

questions. For example, a well-populated database presents opportunities to investigate interactions between populations of different species of arthropod that overlap in geography but may be of interest individually to different realms of research. As a case in point, in the northeastern USA, *Agrilus plannipennis*, the Emerald Ash Borer, is a highly destructive invasive insect, monitored closely by both state and federal agencies for management[60]. Interestingly, Emerald Ash Borers are creating new habitats for carpenter bees, a species interaction that can be tracked and anticipated using large scale arthropod data.

Another example of the utility of linked data is for disease vectors. Data on insecticide resistance linked with time and place would be valuable for coordinating control strategies within and between nations and communities. Presence/absence data on infection levels would be helpful for tracking and investigating disease outbreaks and dynamics. Standardization of these data would be particularly useful for pathogens that infect multiple vectors and hosts and would facilitate a "One Health" approach. Other important vector phenotypes that contribute to control and transmission such as pathogen susceptibility, biting preferences, and breeding behaviours could be measured over time and space.

Indeed, MIReAD would be useable for any arthropod abundance data collection effort, not just medical, veterinary, and agricultural pests or invasive arthropods. We note this standard is applicable not only to abundance measurements, but could be easily extended to any other kind of routinely sampled time-series field data. For example, in addition to aphid abundance, plant pathogen (such as mosaic virus) infection and insecticide resistance statuses of the aphids could be reported in MIReAD format. Note that MIReAD can also be used for cross-sectional data (*i.e. a* non-continuous, one-time sampling effort) by simply reporting data from the single collection period by utilizing a single Start Time and End Time (Online-only Table 2).

Disseminating MIReAD. Many data generators are already storing or sharing data in a manner that would be consistent with MIReAD (*e.g.* on MosquitoNet or NEON), but we call on data generators, authors, reviewers, editors, journals, research infrastructures (*e.g.* data repositories) and funders to embrace MIReAD as a standard to facilitate FAIR data use and compliance for arthropod abundance data. We propose that workshops, outreach at conferences and meetings, and interfacing with data repositories, societies and organizations (*e.g.* SpeciesLink, the American Mosquito Control Association, MosquitoNet, Symbiota, VectorNet, and VectorBiTE), and journal editors will be the best way to spread the adoption of this standard.

Conclusion

We present MIReAD as a minimum information standard for representing arthropod abundance data. MIReAD will facilitate collation and analyses of data at scales that cannot be attained by a single individual or lab in order to address key questions across temporal and spatial scales, such as within and across-year phenology of abundance of target arthropod taxa over large geographical areas. This is particularly important given the pressing need to understand and predict the population dynamics of harmful (*e.g.*, disease vectors and pests) as well as beneficial (*e.g.*, pollinators, bio-control agents) arthropods in natural and human modified landscapes. This is the first step for achieving the broad benefits of FAIR data for arthropod abundance.

Data Availability

No novel data were generated for this report. We encourage readers to view the datasets that inspired and informed our work at www.vectorbase.org, www.gbif.org, www.vectorbyte.org, and in our other publication[14].

References

1. Seastedt, T. R. & Crossley, D. A. The influence of arthropods on ecosystems. *Bioscience* **34**, 157–161 (1984).
2. Moore, J. C., Walter, D. E. & Hunt, H. W. Arthropod regulation of micro- and mesobiota in below-ground detrital food webs. *Annu. Rev. Entomol.* **33**, 419–439 (1988).
3. Whiles, M. R. & Charlton, R. E. The ecological significance of tallgrass prairie arthropods. *Annu. Rev. Entomol.* **51**, 387–412 (2006).
4. Losey, J. E. & Vaughan, M. The economic value of ecological services provided by insects. *Bioscience* **56**, 311–323 (2006).
5. Bradshaw, C. J. A. *et al*. Massive yet grossly underestimated global costs of invasive insects. *Nat. Commun.* **7**, 12986 (2016).
6. Bebber, D. P., Ramotowski, M. A. T. & Gurr, S. J. Crop pests and pathogens move polewards in a warming world. *Nat. Clim. Chang* **3**, 985–988 (2013).
7. Sparling, P. F., Hamburg, M. A., Relman, D. A., Choffnes, E. R. & Mack, A. *Vector-Borne Diseases: Understanding the Environmental, Human Health, and Ecological Connections, Workshop Summary. Forum on Microbial Threats: Board on Global Health*. p. 1–40, (National Academies Press, 2008).
8. Minjauw, B. & McLeod, A. *Tick-borne diseases and poverty: the impact of ticks and tick-borne diseases on the livelihoods of small-scale and marginal livestock owners in India and eastern and southern Africa*. 8, (Centre for Tropical Veterinary Medicine, 2003).
9. Van den Bossche, P., de La Rocque, S., Hendrickx, G. & Bouyer, J. A changing environment and the epidemiology of tsetse-transmitted livestock trypanosomiasis. *Trends Parasitol.* **26**, 236–243 (2010).
10. World Health Organization. *Vector-borne diseases*, http://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases (2017).
11. Golding, N. *et al*. Integrating vector control across diseases. *BMC Med.* **13**, 249 (2015).
12. Elbers, A. R. W., Koenraadt, C. J. M. & Meiswinkel, R. Mosquitoes and *Culicoides* biting midges: vector range and the influence of climate change. *Rev. Sci. Tech.* **34**, 123–137 (2015).
13. Sakai, A. K. *et al*. The population biology of invasive species. *Annu. Rev. Ecol. Syst.* **32**, 305–332 (2001).
14. Rund, S. S. C., Moise, I. K., Beier, J. C. & Martinez, M. E. Rescuing troves of data to tackle emerging mosquito-borne diseases. *J. Am. Mosq. Control Assoc.* **35**, 75–83 (2019).
15. Foley, D. H., Maloney, F. A. Jr., Harrison, F. J., Wilkerson, R. C. & Rueda, L. M. Online spatial database of US Army Public Health Command Region-West mosquito surveillance records: 1947–2009. *US Army Med. Dep. J.* **Jul–Sep**, 29–36 (2011).
16. Hutchinson, M. L., Strohecker, M. D., Simmons, T. W., Kyle, A. D. & Helwig, M. W. Prevalence rates of *Borrelia burgdorferi* (Spirochaetales: Spirochaetaceae), Anaplasma phagocytophilum (Rickettsiales: Anaplasmataceae), and *Babesia microti* (Piroplasmida: Babesiidae) in host-seeking *Ixodes scapularis* (Acari: Ixodidae) from Pennsylvania. *J. Med. Entomol.* **52**, 693–698 (2015).

17. Magarey, R. D. *et al*. Risk maps for targeting exotic plant pest detection programs in the United States: US risk maps for exotic plant pest detection. *EPPO Bulletin* **41**, 46–56 (2011).
18. Wilson, B. E., Beuzelin, J. M., VanWeelden, M. T., Reagan, T. E. & Way, M. O. Monitoring Mexican rice borer (Lepidoptera: Crambidae) populations in sugarcane and rice with conventional and electronic pheromone traps. *J. Econ. Entomol.* **110**, 150–156 (2017).
19. Chandler, M. *et al*. Contribution of citizen science towards international biodiversity monitoring. *Biol. Conserv.* **213**, 280–294 (2017).
20. Kampen, H. *et al*. Approaches to passive mosquito surveillance in the EU. *Parasit. Vectors* **8**, 9 (2015).
21. Suprayitno, N., Narakusumo, R. P., von Rintelen, T., Hendrich, L. & Balke, M. Taxonomy and biogeography without frontiers - WhatsApp, Facebook and smartphone digital photography let citizen scientists in more remote localities step out of the dark. *Biodivers. Data J*. e19938 (2017).
22. Seltmann, K. C. *et al*. LepNet: The Lepidoptera of North America Network. *Zootaxa* **4247**, 73–77 (2017).
23. Short, A. E. Z., Dikow, T. & Moreau, C. S. Entomological collections in the age of Big Data. *Annu. Rev. Entomol.* **63**, 513–530 (2018).
24. Horton, R. (Comment) Offline: What is medicine's 5 sigma? *The Lancet* **235**, 1380 (2015).
25. Nakagawa, S. & Parker, T. H. Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum. *BMC Biol.* **13**, 88 (2015).
26. Nosek, B. A. *et al*. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
27. Parker, T. H. *et al*. Transparency in ecology and evolution: Real problems, real solutions. *Trends Ecol. Evol.* **31**, 711–719 (2016).
28. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci* **3**, 160384 (2016).
29. Ihle, M., Winney, I. S., Krystalli, A. & Croucher, M. Striving for transparent and credible research: Practical guidelines for behavioral ecologists. *Behav. Ecol* **28**, 348–354 (2017).
30. Poisot, T., Mounce, R. & Gravel, D. Moving toward a sustainable ecological science: don't let data go to waste! *Ideas in Ecology and Evolution* **6**, 11–19 (2013).
31. Roche, D. G. *et al*. Troubleshooting public data archiving: Suggestions to increase participation. *PLoS Biol.* **12**, e1001779 (2014).
32. Culley, T. M. The frontier of data discoverability: Why we need to share our data. *Appl. Plant. Sci.* **5**, 1700111 (2017).
33. Gerstner, K. *et al*. Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. *Methods Ecol. Evol* **8**, 777–784 (2017).
34. Ioannidis, J. P. A. *et al*. Repeatability of published microarray gene expression analyses. *Nat. Genet.* **41**, 149–155 (2009).
35. Gilbert, K. J. *et al*. Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Mol. Ecol.* **21**, 4925–4930 (2012).
36. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public data archiving in ecology and evolution: How well are we doing? *PLoS Biol.* **13**, e1002295 (2015).
37. Renaut, S., Budden, A. E., Gravel, D., Poisot, T. & Peres-Neto, P. Management, archiving, and sharing for biologists and the role of research institutions in the technology-oriented age. *Bioscience* **68**, 400–411 (2018).
38. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
39. Wilkinson, M. D. *et al*. A design framework and exemplar metrics for FAIRness. *Sci. Data* **5**, 180118 (2018).
40. Rund, S. S. C. *et al*. Example Minimum Information for Reusable Arthropod Abundance Data (MIReAD) files. *figshare*, https://doi.org/10.6084/m9.figshare.c.4248320 (2019).
41. Taylor, C. F. *et al*. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **25**, 887–893 (2007).
42. Yilmaz, P. *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
43. Lourenço, A. *et al*. Minimum information about a biofilm experiment (MIABiE): standards for reporting experiments and data on sessile microbial communities living at interfaces. *Pathog. Dis* **70**, 250–256 (2014).
44. Brazma, A. *et al*. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
45. Bustin, S. A. *et al*. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* **55**, 611–622 (2009).
46. York, W. S. *et al*. MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology* **24**, 402–406 (2014).
47. Taylor, C. F. *et al*. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896 (2008).
48. Kelly-Hope, L. A. & McKenzie, F. E. The multiplicity of malaria transmission: a review of entomological inoculation rate measurements and methods across sub-Saharan. *Africa. Malaria J* **8**, 19 (2009).
49. Wieczorek, J. *et al*. Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* **7**, e29715 (2012).
50. Giraldo-Calderón, G. I. *et al*. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res* **43**, D707–13 (2015).
51. Benson, D. A. *et al*. GenBank. *Nucleic Acids Res* **41**, D36–42 (2013).
52. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol* **10**, 980 (2003).
53. GBIF: The Global Biodiversity Information Facility. *What is GBIF?*, https://www.gbif.org/what-is-gbif (2018).
54. Heinrich, P. L., Gilbert, E., Cobb, N. S. & Franz, N. Symbiota collections of arthropods network (SCAN): *A data portal built to visualize, manipulate, and export species occurrences*, http://openknowledge.nau.edu/2258/ (2015).
55. Perryman, S. A. M. *et al*. The electronic Rothamsted Archive (e-RA), an online resource for data from the Rothamsted long-term experiments. *Sci. Data* **5**, 180072 (2018).
56. Gossner, M. M. *et al*. A summary of eight traits of Coleoptera, Hemiptera, Orthoptera and Araneae, occurring in grasslands in Germany. *Sci. Data* **2**, 150013 (2015).
57. Hedefalk, F., Svensson, P. & Harrie, L. Spatiotemporal historical datasets at micro-level for geocoded individuals in five Swedish parishes, 1813–1914. *Sci. Data* **4**, 170046 (2017).
58. The American Society for Cell Biology. San Francisco Declaration on Research Assessment, http://www.ascb.org/wp-content/uploads/2017/07/sfdora.pdf (2012).
59. Chavan, V. & Penev, L. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics* **12**(Suppl 15), S2 (2011).
60. Abell, K. J., Bauer, L. S., Duan, J. J. & Van Driesche, R. Long-term monitoring of the introduced emerald ash borer (Coleoptera: Buprestidae) egg parasitoid, *Oobius agrili* (Hymenoptera: Encyrtidae), in Michigan, USA and evaluation of a newly developed monitoring technique. *Biol. Control* **79**, 36–42 (2014).
61. Dunphy, B. M., Rowley, W. A. & Bartholomay, L.C. A taxonomic checklist of the mosquitoes of Iowa. *J. Am. Mosq. Control Assoc*. **30**, 119–121 (2014).
62. Sucaet, Y., Van Hemert, J., Tucker, B. & Bartholomay, L. A web-based relational database for monitoring and analyzing mosquito population dynamics. *J. Med. Entomol.* **45**, 775–784 (2008).
63. Ó Tuama, E., Braak, K. & Remsen, D. *GBIF Metadata Profile – How-to Guide*, https://github.com/gbif/ipt/wiki/GMPHowToGuide (2011).
64. Wieczorek, J., Döring, M., De Giovanni, R., Robertson, T. & Vieglais, D. *Darwin Core Terms: A quick reference guide*, http://rs.tdwg.org/dwc/terms/index.htm (2018).

Acknowledgements

The seeds of this effort were planted in 2016 at a meeting of VectorBiTE, which is a cross-disciplinary research coordination network (RCN) for disease vectors. Samuel S.C. Rund, Matthew Watts, Kurt Vandegrift, Naveed Heydari, Cynthia Lord, Michael Johansson, Samraat Pawar, and Sadie J. Ryan, received travel funding from NIH grant 1R01AI122284-01 and BBSRC grant BB/N013573/1 as part of the joint [NIH-NSF-USDA-BBSRC] Ecology and Evolution of Infectious Diseases program. Samuel S.C. Rund was funded by the Royal Society (NF140517). Rund, Daniel Lawson, Robert M. MacCallum, Sarah A. Kelly, Gloria I. Giraldo-Calderón and Scott J. Emrich were supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201400029C (VectorBase Bioinformatics Resource Center). Kurt Vandegrift was funded by the National Science Foundation Ecology and Evolution of Infectious Diseases program (1619072). Naveed Heydari and Sadie J. Ryan were funded by National Science Foundation (NSF DEB EEID 1518681). Sadie J. Ryan was additionally funded by NIH 1R01AI136035-01, and CDC grant 1U01CK000510-01: Southeastern Regional Center of Excellence in Vector-Borne Diseases: the Gateway Program. This publication was supported by the Cooperative Agreement Number above from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention. Jennifer M. Zaspel was funded by the National Science Foundation Division of Biological Infrastructure (NSF 1561448, NSF 1601957).

Author Contributions

The project was conceptualized by Samuel S.C. Rund, Lauren Cator and Samraat Pawar. The original draft was prepared by Samuel S.C. Rund, Michael A. Johansson, Naveed Heydari, Kurt Vandegrift, Matthew Watts, and Samraat Pawar. All the authors contributed to reviewing and editing the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019