Minimum Information for Reusable Arthropod Abundance Data (MIReAAD)

Myriad: a countless or extremely great number

Samuel Rund*, srund@nd.edu, VectorBase, Department of Biological Science, University of Notre Dame, IN, USA.

Kyle Braak, kyle.braak@gmail.com, Global Biodiversity Information Facility (GBIF) Secretariat, Copenhagen, Denmark

Lauren Cator, l.cator@imperial.ac.uk, Department of Life Sciences, Imperial College London, UK

Kyle Copas, kcopas@gbif.org, Global Biodiversity Information Facility (GBIF) Secretariat, Copenhagen, Denmark

Scott J. Emrich, semrich@utk.edu, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN

Gloria I. Giraldo-Calderón, ggiraldo@nd.edu, VectorBase - Bioinformatics Resource for Invertebrate Vectors of Human Pathogens, Department of Biological Science, University of Notre Dame, IN, USA.

Michael A. Johansson, mjohansson@cdc.gov, Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, 1324 Calle Cañada, San Juan, PR 00920; Department of Epidemiology, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115

Naveed Heydari, naveedheydari@gmail.com, Center for Global Health and Translational Science, State University of New York Upstate Medical University, Syracuse, NY
544 S Vance St, Lakewood, CO, 80226, USA

Donald Hobern, dhobern@gbif.org, Global Biodiversity Information Facility (GBIF) Secretariat, Copenhagen, Denmark

Sarah A. Kelly, s.kelly@imperial.ac.uk, VectorBase, Vector Immunogenomics and Infection Laboratory, Department of Life Sciences, Imperial College London, UK

Daniel Lawson, daniel.lawson@imperial.ac.uk, VectorBase and Vector Immunogenomics and Infection Laboratory, Department of Life Sciences, Imperial College London, UK

Cynthia Lord, clord@ufl.edu, Florida Medical Entomology Lab, University of Florida-IFAS, Vero Beach, FL

Robert M MacCallum, r.maccallum@imperial.ac.uk, VectorBase and Vector Immunogenomics and Infection Laboratory, Department of Life Sciences, Imperial College London, UK

Dominique G. Roche, dominique.roche@mail.mcgill.ca, Institute of Biology, University of Neuchâtel, 2000, Neuchâtel, Switzerland

Sadie J. Ryan, sjryan@ufl.edu, Quantitative Disease Ecology and Conservation Lab, Department of Geography, University of Florida, Gainesville, FL 32601 USA; Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610 USA; College of Life Sciences, University of Kwa-Zulu Natal, Durban, South Africa

Dmitry Schigel, dschigel@gbif.org, Global Biodiversity Information Facility (GBIF) Secretariat, Copenhagen, Denmark

Kurt Vandegrift, kjv1@psu.edu, Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, 16801, PA, USA

Matthew Watts, m.watts@imperial.ac.uk, Department of Life Sciences, Imperial College London, UK

Jennifer M. Zaspel, zaspelj@mpm.edu, Department of Zoology, Milwaukee Public Museum, 800 W Wells Street, Milwaukee, WI, 53233, USA

Samraat Pawar*, s.pawar@imperial.ac.uk, Department of Life Sciences, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, Berkshire SL5 7PY, United Kingdom

* Corresponding authors

Abstract

Arthropods play a dominant role in natural and human-modified terrestrial ecosystem dynamics. Spatially-explicit population time-series are crucial for statistical or mathematical models of these dynamics and assessment of their veterinary, medical, agricultural, and ecological impacts. Arthropod data have been collected world-wide for over a century, but remain scattered and largely inaccessible. With the ever-present and growing threat of arthropod vectors of infectious diseases and pest species, there are enormous amounts of historical and ongoing surveillance. These data are currently reported in a wide variety of formats, typically lacking sufficient metadata to make reuse and re-analysis possible. We present the first minimum information standard for arthropod abundance. 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]. Meanwhile, invasive insects 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 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. Monitoring arthropod abundance (*e.g.* Figure 1) in different disciplines (*e.g.*, biodiversity research, pest-control assessment, vector-borne disease monitoring, or pollination research) uses similar techniques, with similar objectives: to quantify abundance, phenology and geographical ranges of target arthropod species. Despite a growing number of data collections, they are often not reusable, or comparable to similar data, due to a lack of standardization and metadata. In contrast, the advent of the 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[30–33]. This will also lead to substantial public benefit through improved human, animal, plant, and ecosystem health, and reduced economic costs.

A key impediment 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 (MIReAAD) standard for reporting primarily longitudinal (repeated, temporally explicit) field-based collections of arthropods. In the same manner as has been developed in other biological disciplines[40–45], this standard is "minimum" because it defines the necessary minimal information required to understand and reuse a dataset without consulting any further text, materials, or methods[46]. MIReAAD is designed to facilitate data archiving efforts of publishers and field researchers. It is not a data model and therefore does not define controlled vocabularies, or specific field titles, but should be easy to understand, and interpret by the wider scientific community[46].

The minimal standards are 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 (Table 1) includes information for the origin of the data set (*e.g.* study information and licensing for usage). The second component (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. MIReAAD 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.* Figure 2A). By adhering to MIReAAD standards, omissions and ambiguity can be avoided even if the data are shared in different formats (Figure 2B and C). Finally, we identify common problems likely to be encountered across all the MIReAAD 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 is essential for good data discovery and reuse.

Unambiguous dates. Because of country-level differences in date formats, data should be reported with 4 digit years, and months provided alphabetically and not numerically (*e.g.* 4-Jun-2017 or Nov 12, 2015).

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; Table 1).

Language. Once data are ready to be deposited/submitted, all fields and data are 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, non-English field names in an English-language submission.

Examples

Below we provide three examples to illustrate MIReAAD compliant data (linked to Supplemental Data Files 1-4, 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. Example 3 (Data Files 3-4) demonstrates a more complex data collection scenario.

1. *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. [Sup Datasheet 1]. Also see Figure 2B.

2. *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. [Sup Datasheet 2]. This illustrates an example of adult mosquito populations that have been tracked over time and in specific locations. Also see Figure 2C.

3. *Complex trapping data scenario.* 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) [Sup Datasheet 3] . 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 [Sup Datasheet 4].

Discussion

MIReAAD as the path to FAIR data principles

We designed MIReAAD to achieve a balance between standards that are too onerous for data generators and standards that are sufficient to ensure at least minimal reusability[31,40]. Like all minimum standards, MIReAAD only aims at ensuring data 'Reusability'. However, ultimately this will promote the implementation of data models — the explicit definition of data field names, data formats (*e.g.*, for dates and GPS locations), and controlled vocabularies (*e.g.*, the Darwin Core[47]). Data models enable 'Interoperability', and in turn facilitate structured databases, public repositories, and development of data analysis tools[46,48]. Deposition in open databases make data 'Findable' and 'Accessible'[49–51]. MIReAAD compliant data contain sufficient information for established aggregators/databases such as VectorBase and SCAN (Symbiota Collections of Arthropods Network[52]) 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)[47]], and ultimately facilitate data transfer to even more comprehensive biodiversity databases [*e.g.* GBIF, which contains over one billion species occurrence records, from thousands of environmental, ecological, and natural resource investigations, including research on Arthropoda in numerous ecological and monitoring projects, allowing for study of changes and trends in populations.[51]]. Indeed, in Supplemental File 5, we provide an example of the mapping of data fields from this minimum information standard, to DarwinCore and GBIF. In this way, MIReAAD 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. MIReAAD 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. 2014[31]]. Large, deposited data sets can now themselves be standalone, citable "data papers" (*e.g.* [53–55]) 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)[56], offering further opportunities for data generators to gain recognition and publication credit for their work[57]. 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 MIReAAD will provide a foundational pipeline for stipulating archival formats.

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 MIReAAD therefore can 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 MIReAAD applications and extensions

The creation of minimum information standards for these types of databases facilitates analyses of data at the 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 (EAB), is a highly destructive invasive insect, monitored closely by both state and federal agencies for management[58]. Interestingly, EAB are creating lots of new habitat 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.

We note that MIRreAAD 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 MIRreAAD format.

Conclusion

We present MIReAAD, a minimum information standard for representing arthropod abundance data. MIReAAD will facilitate collation and analyses of data at scales that cannot be attained by a single individual or lab, 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. We call on data generators, authors, reviewers, editors, journals, research infrastructures (*e.g.* data repositories) and funders to embrace MIReAAD as a standard to facilitate FAIR data use and compliance for arthropod abundance data.

Table 1. The MIReAAD Study Information (Resource metadata) fields. The information in this table should be included with every data submission, for example by including data in the file header as demonstrated in Data Files 1-4.

Field	Details	Recommendations	Examples

Contact details	A name, person, authority, etc. that may be contacted with enquiries about the data.	Include investigator ORCID(s), email address, website (if institutional) if possible.	Kurt Vandegrift orcid.org/0000-0002-5690-3300 kurtvandegrift@gmail.com State University Agricultural Extension John Smith (jsmith@StateU.edu) www.StateU.edu/AgriculturalExtension/
General description of the experiment/ collection set	A short description of the study objectives, sampling design, and hypotheses. Used to aid in browsing multiple studies. A short title and long form name might be helpful.	Useful things to indicate are: Random sampling or continuous monitoring in fixed locations General time frames and location. General description of where data is from.	"Long term, fixed trapped, municipal surveillance of west Nile vector population in Colorado from 2000-2010" ----- "Pennsylvania *Ixodes scapularis* weekly abundance" Continuous (weekly) monitoring of tick numbers attached to White-footed mice in fixed locations in Pennsylvania, USA (12 sites). 2003-present." ----- "Long term aphid emergence monitoring using continuous suction traps"
Citations	Reference to related publications, digital if possible (e.g. DOI(s) or PMID(s)).		"A web-based relational database for monitoring and analyzing mosquito population dynamics Sucaet Y, Van Hemert J, Tucker B, Bartholomay L." "PMID: 18714883" Horiuchi, Kaho, Kosei Hashimoto, and Fumio Hayashi. "Cantharidin world in air: Spatiotemporal distributions of flying canthariphilous insects in the forest interior." Entomological Science (2018).
Species Identification Method	A description of method of species identification. Particularly important for cryptic species complexes.		"Morphological" "Genotyped, using method of Smith et al 2014, PMID: 18714883"
Not present vs zero information	Indication of what gaps, zeros, NA, etc mean.	It is imperative, especially for population surveys, to understand the difference between a species was not found when the collection method would be expected to find the given species (confirmed absence) or a species was not	"Zero indicates was looked for and not found. NA represents a trap failure etc"

		looked for (*e.g.* a trap failure) Preferably, a zero indicates was looked for and not found, and a NA represents was not looked for/trap failure/ etc. Blank values are discouraged	
GPS obfuscation information	If GPS data obfuscation (*e.g.* GPS points are intentionally offset from their actual locations) or de-resolution occurs (*e.g.* GPS precision is intentionally reduced) , a statement on the manner by which this occurred.	The highest resolution data (*e.g.* trap-level, specific GPS location) are the most useful. It is hoped that no data obfuscation / de-resolution occurs	"GPS locations have been truncated to 3 decimals" "GPS locations obfuscated using N-Dispersion" "No GPS deresolution was performed"
Data usage information	The data reuse policy for your data. Please provide a creative commons license identification. See https://creativecommons.org for more information.	For data to be F.A.I.R., it must be Reusable. We therefore recommend data be provided as "CC0" or "CC BY 4.0". "CC0", under which data are made available for any use without restriction or particular requirements on the part of users "CC BY 4.0", under which data are made available for any use provided that attribution is appropriately given for the sources of data used, in the manner specified by the owner (*e.g.* citation).	"CC0" or "CC BY 4.0"

Table 2. The MIReAAD data fields. Fig 1B provides an annotated example**.**

Field(s)	Details	Recommendations	Examples
Start Time (for collection)	Start time of the data sample collection. *e.g.* The trap was set...	Be as specific as *practically* possible. Any unambiguous format is acceptable. However, do not use two-digit year abbreviations. If relevant, provide timezone in field or in header, a 24 hour clock is preferred, but should be made unambiguous as to which time format is being used.	"2012-04-27" "July 26, 2017" "2017-Jul-26" "2017-July-26 Morning " "2017-Jul-26 20:00 GMT "
End Time (for collection)	End time of the data sample collection. *e.g.* The trap was collected...	See above. If instantaneous data collection (*e.g.* a tick drag), End Time may be the same as Start Time.	See above.
Location	The geographical location of sample collection.	As detailed as possible. Latitude and longitude if possible with specified accuracy Providing *both* a GPS point (decimalized GPS points are prefered) field and a geographical name field is prefered. Note only providing location *names* is highly discouraged as they change over time and can be ambiguous. Place / Trap names and GPS fields can be provided. If obfuscation was used, it should be indicated in the Metadata (Table 1). Splitting latitude and longitude further into two columns further reduces ambiguity.	"Kukar Maikiya, Jigawa State, Nigeria" "40.697" and " -74.015"

Collection method	Sampling apparatus (*e.g.* trap type, observation method)		"CDC light trap" "Tick drag" "Quadrat count" "BG Sentinel Trap" "Pitfall trap" "Larval dip" "Johnson suction trap" "Lindgren Funnel Trap"
Collection attractants	The attractant/ lures used to attract insects to a trap or collection		"None" "Carbon dioxide" "UV light" "BG-Sweetscent Mosquito Lure" "Human/animal bait"
Collection area	The spatial extent (area or volume) of the sample.	If relevant (*e.g.*, when collection method is transect or quadrat), in units of area or volume, the spatial coverage of the sampling unit Note this field would not typically be used for mosquito collections.	"100 m^2" "1 liter" "1 ha" "10m^3"
Taxonomy	Classification of sample collected.	Scientific genus and species preferred. Avoid abbreviation.	*"Ixodes scapularis*" *"Aedes aegypti*" '*Anopheles gambiae sensu stricto*"

Unit(s) of measurement and observation	Description of exactly what was observed, the unit for "Value" below. For counts, should indicate life stage, sex, etc. Unit measures can be encoded into value field header. Consider multiple unit fields (e.g. separate fields for sex and stage.) See Figure 2.	Do not abbreviate. Coded data key should be provided in field name (*e.g.* "1 = species present 0= species absent")	"Number of individuals per m^2" "Female" and "Adult" "Male and Female" and "Nymphs"
Value	The numerical amount or result from the sample collection. Often this will be a quantity of observed individuals. Unit measures can be encoded into value field header. See Figure 2.	Units should be provided in a separate field.	"0" "23" "Yes" "Not present"
Additional sample information	This could be more than one field and should be used when more information is required to understand the experiment, for example experimental variables, sub-locations, etc. Some users may report wind speeds, temperatures, elevations etc.	Do not abbreviate.	"Forest" vs "Field" "Winter" vs "Summer" "Inside" vs "Outside" "200 meters above sea level"

Sample Name	A human readable sample name. May exist solely for the benefit of the depositor in organizing their data, use their own internal naming conventions etc. May also be used to tie related observations together.	Naming convention is not restricted, but any encoded metadata should be revealed in the other datafields. For example, you may name a sample named 'Aphid1_StickyTrap_Jan4,' but you will still have "Sticky Trap" listed in a Collection Method field, and "Jan 4, 2017" in the date field.	"Trap1_Night1" "KissingBug_2" "00004" "Jan08_animal_4,"

Field names in bold should be considered also required. Remaining fields are optional or depend on the complexity of the experimental design

Author contributions

The project was conceptualized by Lauren Cator and Samraat Pawar. The original draft was prepared by Michael A. Johansson, Samuel S.C. Rund, Naveed Heydari, Kurt Vandegrift, Matthew Watts, and Samraat Pawar. Visualization was prepared by Kurt Vandegrift, Samuel S.C. Rund, Samraat Pawar, and Michael A. Johansson. Review & Editing was performed by all the authors.

Competing interest statement

The authors declare no competing interests.

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-Calderon 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).

Figures

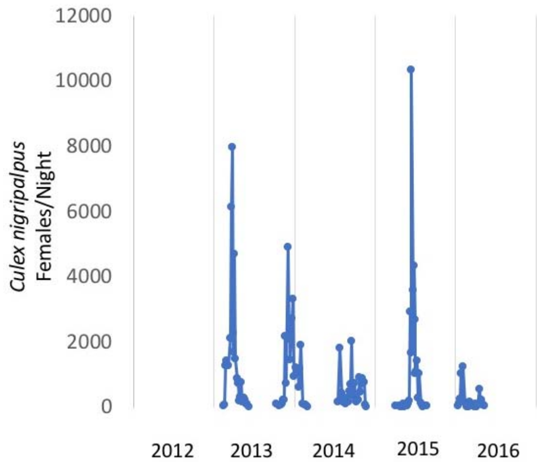

Figure 1. **Example population abundance time-series.**

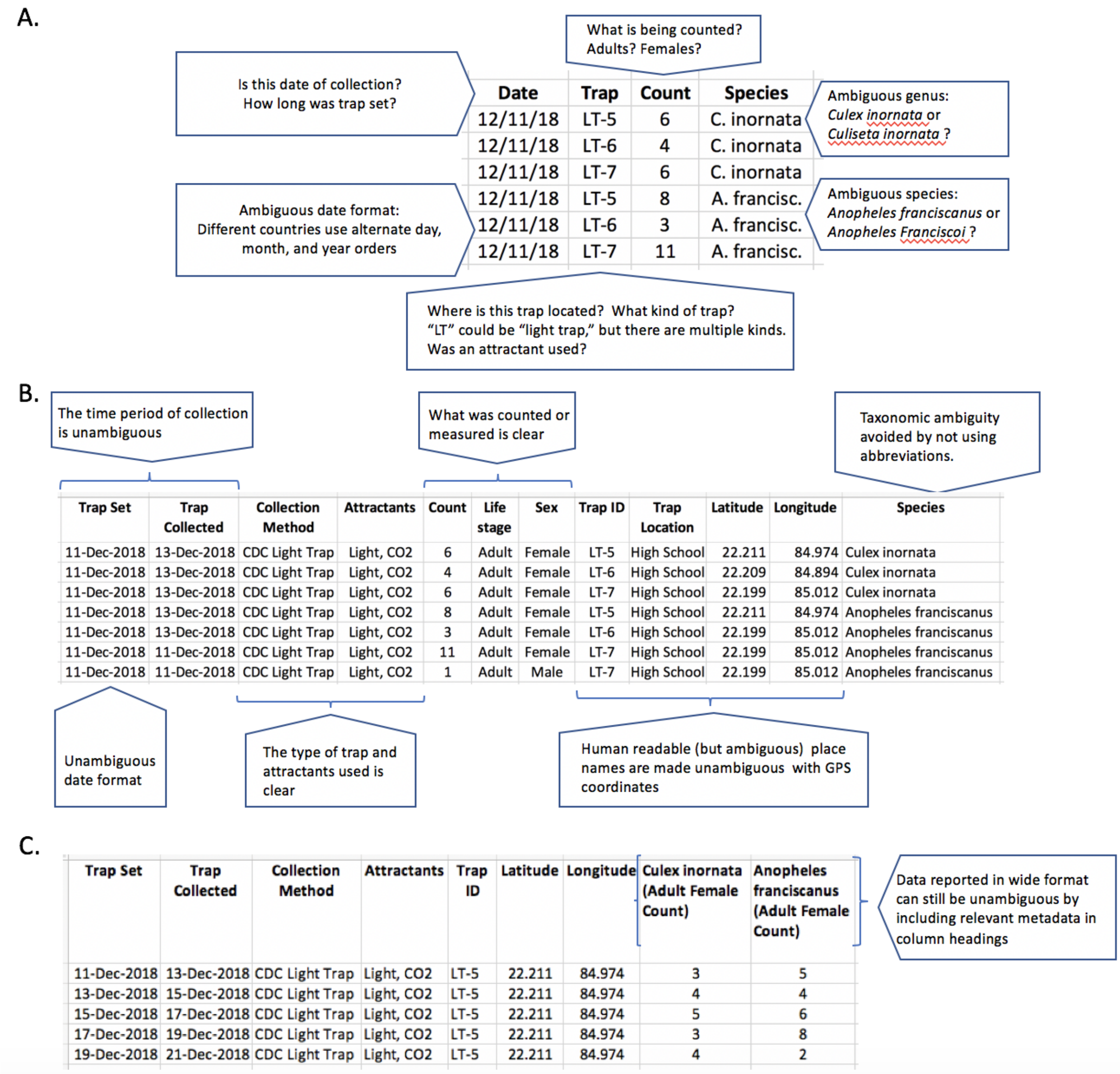

Figure 2. **MIReAAD reduces data ambiguity.** A. Seemingly clean data can still lack key information or have ambiguous metadata, hindering data reuse. B. MIReAAD compliant data includes the metadata necessary for data reuse and removes ambiguity. C. Note data can be formatted differently, but still be MIReAAD complaint, such as by presenting data in a wide format

References

1. Seastedt, T. R. & Crossley, D. A. The Influence of Arthropods on Ecosystems. *Bioscience* **34,** 157–161 (1984).
2. Arthropod Regulation of Micro- and Mesobiota in Below-Ground Detrital Food Webs |

Annual Review of Entomology.
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 (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.* (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*. (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. WHO | Vector-borne diseases. (2017).
11. Gubler, D. J. Resurgent vector-borne diseases as a global health problem. *Emerg. Infect. Dis.* **4,** 442–450 (1998).
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. & Martinez, M. E. Rescuing Troves of Data to Tackle Emerging Mosquito-Borne Diseases. *bioRxiv* 096875 (2018). doi:10.1101/096875
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.* 29–36 (2011).
16. Hutchinson, M. L., STROHECKER, 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. *Journal of Medical Entomology* **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. E., Mounce, R., Gravel - Ideas in Ecology and, D. & 2013. Moving toward a sustainable ecological science: don't let data go to waste! *queens.scholarsportal.info* (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. *Applications in Plant Sciences* **5,** (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. *Wiley Online Library* (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. Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE).

Nat. Biotechnol. **25,** 887–893 (2007).

41. 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).
42. 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).
43. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock - Nature …, G. & 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *nature.com* (2001).
44. Bustin, S. A. *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* **55,** 611–622 (2009).
45. York, W. S. *et al.* MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology* **24,** 402–406 (2014).
46. Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26,** 889–896 (2008).
47. Wieczorek, J. *et al.* Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* **7,** e29715 (2012).
48. 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).
49. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41,** D36–42 (2013).
50. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10,** 980 (2003).
51. GBIF. Available at: http://gbif.org. (Accessed: 26th March 2018)
52. 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.
53. 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).
54. 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).
55. 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).
56. The American Society for Cell Biology. San Francisco Declaration on Research Assessment. (2012). Available at: http://www.ascb.org/wp-content/uploads/2017/07/sfdora.pdf.
57. Chavan, V. & Penev, L. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics* **12 Suppl 15,** S2 (2011).
58. 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).