*Journal of Medical Entomology*, XX(XX), 2023, 1–8
https://doi.org/10.1093/jme/tjad009
Forum

ENTOMOLOGICAL SOCIETY OF AMERICA
SHARING INSECT SCIENCE GLOBALLY

OXFORD

Forum

# Characterizing the Vector Data Ecosystem

**Catherine A. Lippi,[1,2] Samuel S. C. Rund,[3] and Sadie J. Ryan[1,2,4]**

[1]Quantitative Disease Ecology and Conservation (QDEC) Lab Group, Department of Geography, University of Florida, Gainesville, FL 32611, USA, [2]Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610, USA, [3]Center for Research Computing, Department of Biological Sciences, & Eck Institute for Global HealthUniversity of Notre Dame, Notre Dame, IN 46556, USA, and [4]Corresponding author, e-mail: sjryan@ufl.edu

Subject Editor: Ary Faraji

Received 29 September 2022; Editorial decision 3 January 2023.

## Abstract

A growing body of information on vector-borne diseases has arisen as increasing research focus has been directed towards the need for anticipating risk, optimizing surveillance, and understanding the fundamental biology of vector-borne diseases to direct control and mitigation efforts. The scope and scale of this information, in the form of data, comprising database efforts, data storage, and serving approaches, means that it is distributed across many formats and data types. Data ranges from collections records to molecular characterization, geospatial data to interactions of vectors and traits, infection experiments to field trials. New initiatives arise, often spanning the effort traditionally siloed in specific research disciplines, and other efforts wane, perhaps in response to funding declines, different research directions, or lack of sustained interest. Thusly, the world of vector data – the Vector Data Ecosystem – can become unclear in scope, and the flows of data through these various efforts can become stymied by obsolescence, or simply by gaps in access and interoperability. As increasing attention is paid to creating FAIR (Findable Accessible Interoperable, and Reusable) data, simply characterizing what is 'out there', and how these existing data aggregation and collection efforts interact, or interoperate with each other, is a useful exercise. This study presents a snapshot of current vector data efforts, reporting on level of accessibility, and commenting on interoperability using an illustration to track a specimen through the data ecosystem to understand where it occurs for the database efforts anticipated to describe it (or parts of its extended specimen data).

**Key words:** vector-borne disease, mosquito, database, interoperability, ecoinformatics

Vector-borne diseases pose a major threat to public health and agricultural systems globally (Institute of Medicine (US) Forum on Microbial Threats 2008, Golding et al. 2015, Kitsou and Pal 2022). These systems are typically complex and span multiple spatial and temporal scales. Consequently, there is often a considerable burden of data needed to conduct meaningful research in this area (Cator et al. 2020). Repositories that aggregate disease vector data from multiple sources (e.g., museums, individual research projects, or public health surveillance systems) broaden the horizon of research possibilities, in some cases alleviating the logistical constraints of novel data collection (Suarez and Tsutsui 2004, Kampen et al. 2015, Trivellone et al. 2021). The goal of this piece is to document the state of the 'ecosystem' of vector databases that we and other researchers in vector-borne disease systems may contribute to, use, and often reuse and repurpose. For the sake of simplicity, we use the term vector to denote the organism that is the route of indirect transmission for an infection or pathogen, across animal and plant systems. While this will largely be represented by arthropod insects, with a focus on vectors of human disease, we wanted to include as wide a breadth as possible.

One aspect of database use, reuse, production, and augmentation is access, which can be limited to a level less than entirely open due to a multiplicity of factors. The issue of privacy of specific records can be tied to the same restrictions on identifiability of human subjects; for example if an infected vector is located at a household with a particular suite of demographic descriptors, the threshold for identifiability may be an issue (Secunda 2004, Moy et al. 2018). Similarly, if infected crop pests are identified at a particular location, this may violate privacy of an agricultural business enterprise. Beyond privacy for identifiability, ownership of data can be subject

to contributor agreements and the data may be generated using resources with specific intellectual property or ownership stipulations (Scheibner et al. 2021). These first limits to access are generally well understood and accepted, and working to create derived data products, such as data summaries, de-identified data, or explicit data use agreements, can start to lower the barriers to data access in these scenarios. Additional limits to access may include limited resources to hosting, data management, digitization, formatting, creating access portals, and other informatics related issues (National Research Council (US) Board on Biology 2010, de Carvalho Gomes et al. 2021). In some cases, the onward utility of a database, generated for a specific project use, or a specific ongoing purpose (e.g., monitoring crop pests, vector surveillance), may not be realized as part of the value of the product or project goals. Thus these datasets may remain siloed, stored on a single computer for a bespoke purpose, even finished and shelved (or lost, corrupted, deleted, or destroyed). This latter set of scenarios is less well appreciated, and constraints may include the database construction or data entry itself, and the knowledge that a dataset even exists.

In previous efforts, we have attempted to address part of the latter access limit scenarios, though outreach to nonacademic groups collating and curating vector records (Rund et al. 2019b); through creating explicit data structure and data entry primers (Rund et al. 2019); and hosting workshops for data users to think through the processes of reconciling disparate data sources and data sets, and using the feedback to recycle back into outreach and information. As an additional step to addressing some of the access limits we identified, we felt it would be useful to create a (nonexhaustive) compendium of sources, providing:

1. A short description of intent and scope (e.g. data type, taxonomic, geographic, and temporal)
2. Current repository or effort location and access point
3. Accessibility (i.e., fully accessible versus partial or limited accessibility)

## Fully Accessible Databases

### Global Biological Information Facility (GBIF) and Aggregators

GBIF (https://www.gbif.org) is a world-wide index for species occurrence records from across the tree of life, with over 1 billion species occurrence records. Although GBIF is not dedicated to hosting entomological surveillance data, the index is nevertheless an extensive data source for georeferenced vector records (e.g., there are over 1.8 million mosquito occurrence records on GBIF as of October 2022). This comprehensive index relies on a network of partners, many specific to taxa or geographic region, that aggregate and feed data into GBIF. GBIF's formal partners establish 'nodes', or teams designated to coordinate and manage the flow of biodiversity data, according to geographic regions or themes (GBIF Secretariat 2020). The primary aggregator that serves data to GBIF from the United States is the United States Geological Survey (USGS) Biodiversity Serving Our Nation (BISON) program (https://bison.usgs.gov/ipt/). BISON (USGS 2013) is a federal mapping resource for species occurrence data and contributes to the US Node of GBIF, focusing on government collections and invasive species in the United States, U.S. associated territories, and Canada.

There are many other entities that contribute vector data to GBIF. The National Science Foundation (NSF) National Ecological Observatory Network (NEON, https://www.neonscience.org) monitors ecosystems across the United States, providing time series and abundance data for species (including vectors) across the project's field sites, which include 47 terrestrial sites. Integrated Digitized Biocollections (iDigBio, https://www.idigbio.org) is an initiative to digitize museum holdings undertaken by the National Resource for Advancing Digitization of Biodiversity Collections (ADBC) and funded by the NSF. VertNet (http://www.vertnet.org) is another NSF-funded collaboration to streamline the availability of vertebrate biodiversity data, which may include arthropods associated with records (e.g., parasites). Though not a data provider, GBIF has also partnered with the GigaByte journal, a publishing platform that supports data releases, to publish new datasets in the 'Vectors of Human Disease Series', a thematic data release series that is also available on GBIF (Gigabyte: Vectors of human disease series 2022).

Large community science databases also provide data to GBIF. iNaturalist (https://www.inaturalist.org), a joint initiative between the California Academy of Sciences and the National Geographic Society, solicits observations from the public that are identified by users on the platform. While the primary purpose of iNaturalist is to promote community engagement with nature, the online community also generates a great deal of georeferenced occurrence data. Data fed into GBIF meet the iNaturalist criteria for 'research-grade' observations, where occurrences typically have a photo, geographic coordinates, and community consensus on identification. Similarly, BugGuide (https://bugguide.net) is a platform hosted by Iowa State University Department of Entomology, and provides community science occurrence data focused on insects, spiders, and related arthropods. It must be noted that some of these databases may only share a subset of their data with GBIF. For example, iNaturalist contains many citizen-science occurrence records, but these are not all shared as they are not considered 'research-grade'.

Another aggregator of note is the Symbiota Collections of Arthropods Network (SCAN) (https://scan-bugs.org/). They serve as a regional GBIF node specializing in providing arthropod occurrence data, aggregating records from over 225 data providers in North America. Providers include collections maintained by academic institutions, natural history museums, government agencies, and more. Collections that share data through SCAN vary considerably in focus, ranging from general entomology to specialized collections, such as medically important arthropods or agricultural pests. Although SCAN has a primarily North American focus, the data they provide is global in scope. This is the node that VectorBase uses to share data with GBIF. SCAN also aggregates arthropod occurrence data from BISON, NEON, iDigBio, iNaturalist, BugGuide, VertNet, the Terrestrial-Parasite-Tracker Thematic Collection Network (TCN), and others. All data on GBIF are publicly available.

### TPT (Terrestrial Parasite Tracker) TCN (Thematic Collection Network)

The Terrestrial Parasite Tracker (TPT) is a new project funded by the NSF's ADBC program to facilitate the digitization of arthropod ectoparasite and vector specimens held in natural history museums collections (Poelen et al. 2021). Additionally, the TPT aims to offer support and resources for the digitization of 'hidden' collections, or holdings not associated with institutional collections that are not within the purview of the iDigBio project. In addition to digitization of physical specimens and georeferenced locality data, special emphasis is placed on capturing host:parasite interaction/relationship data. An overarching goal of the TPT project is to integrate arthropod ectoparasite data into GBIF. The TPT data are associated with the Global Biotic Interactions (GloBI) portal (https://www.globalbioticinteractions.org/parasitetracker) (Poelen et al. 2014), and accessible through SCAN.

### VectorBase

VectorBase (https://vectorbase.org) (Giraldo-Calderón et al. 2021) has been in existence for over 17 years. VectorBase is primarily known as an extensive genomics data resource, with hundreds of datasets deposited spanning genome assemblies, proteomics, gene expression studies, population genetics, and data on genetically based phenotypes (e.g., insecticide resistance in vectors). In addition to molecular data, VectorBase has over 1,600,000 nonzero abundance records and over 25 million collection events that did not detect vectors. Data providers typically share surveillance data with VectorBase as one-off or yearly data exchanges that are manually processed by VectorBase staff, or data is curated directly from published literature. In regards to U.S. vector population surveillance data, these data may come from a U.S. state level or from local authorities. Particularly rich data come from Florida, where numerous programs submit data.

Field-collected spatial and temporal data on arthropod vectors are available to browse via a custom-built web application, the MapVEu system (https://vectorbase.org/popbio-map/web/). MapVEu facilitates map-based data exploration with location and metadata search features that drive dynamically generated mapping, live graphing, and interactive display of data. This specialized mapping interface enables viewing of a variety of data including vector population abundance surveys, insecticide resistance genotypes and phenotypes, blood meal host analysis, and pathogen testing results.

In 2019 VectorBase was merged with the Eukaryotic Pathogen Genomics Database Resource (EuPathDB) to form the Eukaryotic Pathogen, Vector and Host Informatics Resource (Amos et al. 2022) (VEuPathDB). As part of VEuPathDB, VectorBase remains a distinct database, but utilizes a shared web infrastructure within the VEuPathDB project. Collectively, these resources comprise one of the two Bioinformatics Resource Centers (BRCs) for infectious diseases supported by the US National Institute of Allergy and Infectious Diseases (NIAID) and National Institutes of Health (NIH) (https://www.niaid.nih.gov/research/bioinformatics-resource-centers). All data on VectorBase are publicly available.

### VectorBiTE/VectorByte

Vector Behavior in Transmission Ecology (VectorBiTE) (https://www.vectorbite.org) is a research coordination network (RCN) co-funded by US National Institutes of Health (NIH) and UK Biotechnology and Biological Sciences Research Council (BBSRC). The primary focus of the network is to facilitate collaborations among the diverse fields studying vector-borne diseases, thus promoting the advancement of vector modeling research. In addition to fostering collaborations, VectorBiTE has several key objectives, which include improved data collection standards, statistical methods, and the development of validation datasets to aid in model development and comparison. These objectives have already yielded results, such as the minimum information standards outlined by the Minimum Information for Reusable Arthropod Abundance Data (MIReAAD) (Rund et al. 2019a), a paper that was developed by collaborators in the VectorBiTE consortium.

VectorByte (https://www.vectorbyte.org), the successor project to the VectorBiTE RCN, began in August 2020. The goal of the VectorByte initiative is to establish a global, open access data platform to support research on vector-borne diseases. The VectorByte data hub comprises two separate databases, VecTraits and VecDyn. VecTraits hosts curated ecological trait data for vectors and some pathogens, such as temperature-dependent growth and survival rates, fecundity, and vector competence. VecDyn is a population abundance database, conducive to supporting research on vector population dynamics. VecDyn incorporates data from sources involved in long term vector research. While it primarily hosts mosquito data aggregated from other databases, the database can also host nonhuman vectors of livestock and plants. All data available through VectorByte are publicly available.

### VectorMap

The VectorMap Data Portal (https://vectormap.si.edu) is a product of the Walter Reed Biosystematics Unit (WRBU), a partnership between the Walter Reed Army Institute of Research (WRAIR), and the Smithsonian Institution National Museum of Natural History (NMNH). VectorMap holds vast amounts of well curated, high-confidence, geospatial species occurrence data for a wide variety of medically important arthropod taxa (Foley et al. 2009, 2010), including mosquitoes (MosquitoMap), ticks (TickMap), fleas (FleaMap), mites (MiteMap), biting midges (MidgeMap), and sandflies (SandFlyMap). VectorMap has approximately 700,000 records of vector surveillance data. Database records are routinely added to the database from a variety of sources, which include military biosurveillance initiatives, ongoing entomological surveillance, digitization of museum collections, and datasets published in scientific literature. The VectorMap interface allows users to interactively browse and view records through dynamic mapping and search functions. In addition to occurrence records for medically important arthropods, VectorMap also holds data on blood meal analysis, hosts, and insecticide resistance, as well as exportable niche models of habitat suitability for select vector species. Vector Hazard Reports (VHRs) are another product available through the data portal, where VectorMap data are combined with other risk indicators to produce risk profiles for discrete countries or regions (Walter Reed Biosystematics Unit 2022). All data on VectorMap are publicly available.

To streamline the submission of entomological surveillance data, VectorMap released a best practices guide for data formatting and curation. This guide includes recommendations for formatting locality data and minimum reporting standards (Walter Reed Biosystematics Unit 2021), aiding in post hoc georeferencing and increasing broader utility and usability of datasets.

## Partially Accessible Databases

**CDC ArboNET** The National Arbovirus Surveillance System (ArboNET, https://wwwn.cdc.gov/arbonet/maps/ADB_Diseases_Map/index.html) is managed by the US Centers for Disease Control and Prevention (CDC) in cooperation with state health departments (https://wwwn.cdc.gov/arbonet/). The ArboNET system relies on passive surveillance, such as clinician diagnosis, testing, and reporting to local public health authorities. Reported data include human arboviral disease cases, and non-human infections from mosquito populations, veterinary cases, wildlife, and sentinel surveillance animals. The ArboNET system is provided for government-authorized use only, and users may request an account through the data portal. However, data aggregated to the county level are viewable through the CDC ArboNET Disease Maps website. Data available through the mapping platform include aggregated human arbovirus cases, neuroinvasive disease incidence, and locally acquired versus imported cases of dengue, chikungunya, and Zika. In addition to human cases, the presence of county-level infections (i.e., presence/absence data) from veterinary, sentinel animal, avian, and mosquito infections are visible for a number of established arboviruses. Data hosted on CDC ArboNET are partially accessible, as aggregated summary data are viewable online.

**Ecological Database of the World's Insect Pathogens (EDWIP)** The Ecological Database of the World's Insect Pathogens (EDWIP) database contains associations of pathogens with insects and other arthropods. The database was largely compiled in the 1990's, and was first described in Braxton et al. (2003). The EDWIP is notable in that, in addition to naturally occurring host-parasite relationships, it also includes some data that were derived from experiments where hosts (e.g., insects) were inoculated with pathogens but did not become infected (Braxton et al. 2003). Thus it contains some 'true absences' in the documented associations. This database also includes some ecological data associated with hosts and parasites (e.g., habitat and diet of hosts). In 2015, as part of an effort by the NSF funded Macroecology of Infectious Diseases RCN, a large portion (~3,000 rows) of the database was made into an easy-to-read CSV file, with documentation recorded online (https://edwip.ecology.uga.edu/). In 2021, the R package insectDisease was created and stored on GitHub, to facilitate access to data and documentation (https://github.com/viralemergence/insectDisease). Although the database is available through the R package, data downloaded through the website may not be complete, and these two resources are not reconciled as of October 2022.

**IR Mapper** IR Mapper (https://www.irmapper.com) is an online, interactive mapping tool that displays insecticide resistance testing data for Anopheles species, and two arboviral vectors, *Aedes aegypti* (Linnaeus, 1762) (Diptera: Culicidae) and *Aedes albopictus* (Skuse, 1895) (Diptera: Culicidae). Started in 2012, the platform hosts data generated using CDC or World Health Organization (WHO) testing protocols for resistant phenotypes and genotypic resistance mechanisms. Data are viewable through interactive mapping functions, and are mostly obtained on a monthly basis from peer-reviewed published literature, although other sources of insecticide resistance data are also used, such as published reports. While there are functions to export mapped results of queries, data for *Aedes* records are accessible indirectly via cited literature for individual records, and the *Anopheles* mapping interface now has an option for direct data downloads (Moyes et al. 2019). IR Mapper is a joint initiative between Vestergaard and the Kenyan Medical Research Institute Centre for Global Health Research.

**Malaria Atlas Project** The Malaria Atlas Project (MAP) (https://malariaatlas.org) is an online platform founded in 2005 that hosts interactive mapping, trend visualization tools, and data directories for malaria and associated mosquito vectors. Primarily funded by the Bill & Melinda Gates Foundation, MAP collaborates with the WHO, and has been designated as a WHO Collaborating Centre in Geospatial Disease Modelling. Data on vector occurrence, malaria prevalence, and covariates are generally available as spatial layers, downloadable through the platform's Data Explorer mapping interface. Model outputs of risk and predicted geographic vector ranges are also available through this platform as layers (Hay and Snow 2006). Although many datasets hosted by MAP are openly available, accessibility and permissions vary across datasets.

**Malaria Threat Map** Malaria Threat Map (https://apps.who.int/malaria/maps/threats) is an interactive data and mapping platform produced by the WHO. This database specializes in biological challenges to malaria control and elimination, such as vector insecticide resistance and parasite drug resistance. For Anopheline vectors, insecticide resistance phenotype data based on WHO assays and maps of invasive malaria vector occurrence are viewable. Data on malaria parasites include the resistance to the drug artemisinin, a core antimalarial compound, and *pfhrp2* gene deletions, which cause false negatives in rapid diagnostic tests (RDTs) for malaria (Koita et al. 2012). Data on Malaria Threat Map are selected via filter options in the online mapping platform. Data are downloadable after completion of an online form, where users provide contact information, professional affiliation, and a detailed description of intended data use. The availability of data varies based on permissions established by individual data contributors (e.g., member states, research institutions, scientific publications), and therefore not all data are available for download.

**VectorSurv State Repositories** The Vectorborne Disease Surveillance System (VectorSurv, https://vectorsurv.org) is the umbrella name for a family of state or territory specific web services for vector control and public health agencies in the United States, and U.S.-affiliated Pacific islands. This surveillance network, initially limited to California (CalSurv), began in 2006 as a partnership between public health vector control entities in the state, including the Mosquito and Vector Control Association of California, the California Department of Public Health, and the University of California Davis Arbovirus Research and Training (DART) Laboratory (Barker et al. 2010). In 2017 VectorSurv expanded beyond California, and now includes the partner states of Arizona, California, Hawaii, Nebraska, New Jersey, North Carolina, North Dakota, South Dakota, Tennessee, Utah, Washington, and the U.S.-Affiliated Pacific Islands of Guam, Commonwealth of the Northern Mariana Islands, Federated States of Micronesia, Republic of Palau, and Republic of the Marshall Islands.

VectorSurv is a powerful system for reporting data on abundance and pathogen testing for mosquitoes, ticks, or other arthropod vectors, serological surveillance in sentinel chickens, insecticide resistance testing, and public-health pesticide applications. The platform is unique in that it is designed for day-to-day operational data entry, as opposed to submitting processed data at the end of surveillance season or post-publication. It has numerous tools for analyzing and reporting data that would be helpful to abatement districts generating surveillance data. These include an interactive mapping interface for viewing surveillance data, VectorSurv Maps (https://maps.vectorsurv.org), and the VectorSurv Gateway (https://gateway.vectorsurv.org), an online portal that offers management solutions for facilitating data entry, geospatial analyses, mosquito pool testing for viruses, and calculators to estimate arboviral risk. Data requests can be made to VectorSurv, or directly to any of its partner agencies, and are approved on a case-by-case basis. Although the VectorSurv database is not openly accessible, arboviral mosquito surveillance and partially available data, such as sentinel animal data, are freely viewable through VectorSurv Maps.

**VectorNet** The European Network for Medical and Veterinary Entomology (VectorNet) (https://vectornet.ecdc.europa.eu) is a joint initiative of the European Food Safety Authority (EFSA) and the European Centre for Disease Prevention and Control (ECDC). The Project supports the collection of data on vectors and pathogens in vectors, related to both animal and human health (Braks et al. 2022). The database is closed access, but maps of surveillance efforts and mosquito distributions based on surveillance data are available online. Mapped vector distributions throughout Europe and neighboring regions are available for a number of vectors, including mosquitoes, ticks, phlebotomine sandflies, and biting midges.

## Other Databases

This review has described major repositories of arthropod vector data currently online. However, this list is certainly not exhaustive as

a number of additional bespoke databases exist, often specialized for a targeted audience of users, or a specific analytical purpose. Here, we describe several additional vector databases or datasets of note, which are not necessarily connected to the broader data ecosystem.

Mosquito surveillance data may be accessible directly from abatement programs or research projects. One of the largest accessible platforms for this type of data is maintained by the State of Iowa, which provides a centralized database of mosquito surveillance that is available online (https://mosquito.ent.iastate.edu) through a partnership between the Iowa State University Medical Entomology Laboratory and the Iowa Department of Public Health. Mosquito surveillance data, including mosquito population abundance data for a wide variety of species, are available from 1969 through 2021 at the time of writing (Sucaet et al. 2008). These surveillance data are openly accessible, though geographically restricted to Iowa. VectorMap-GR (https://vectormap-gr.com/) is a similar geographically-restricted database of mosquito populations, limited to Crete (Fotakis et al. 2021). It also includes such data as confirmed larval habitats and insecticide resistance assays. MosquitoDB (https://mosquitodb.io/mdb/login.php) is an African-led project to collate mosquito data, primarily from national malaria control programs maintained by the Pan-African Mosquito Control Association. At this time, it is closed access. Its data model is based on previously published work from the Ifakara Health Institute, Tanzania (Kiware et al. 2016). ClinEpiDB (https://clinepidb.org) curates epidemiological data from large (human) field trials - some of which have paired vector data (Ruhamyankaka et al. 2019).

Beyond formal surveillance systems, there are efforts to produce new and accessible datastreams, adopting novel technologies and reporting chains. Mosquito Alert (http://www.mosquitoalert.com/en/) (Delacour-Estrella et al. 2014) is a nonprofit citizen science project, whereby the public submits pictures of mosquitoes and larval sites using a mobile phone app. Data are openly accessible online and available for download through the Mosquito Alert Data Portal. The Global Learning and Observations to Benefit the Environment (GLOBE) program (https://observer.globe.gov), sponsored by the National Aeronautics and Space Administration (NASA), is a community science application that invites users to submit environmental observations. GLOBE's Mosquito Habitat Mapper provides a tool for users to submit observations on potential mosquito breeding habitat, which may also include larval mosquito presence (Low et al. 2021). Tick Report (https://www.tickreport.com/stats) is a commercial testing service to detect pathogens in user-submitted tick samples. One of the few examples of a vector database which does not exclusively focus on mosquitoes, Tick Report makes data and summary statistics from their testing program available online.

Mosquito surveillance data, which may capture presence and abundance, can be leveraged for a wide range of modeling applications. However, novel analytic techniques may require specific data inputs that are not routinely captured in existing databases. For example, WingBank (https://wingbank.butantan.gov.br) is a database of over 10,000 images of mosquito wings that could have applications for AI-driven mosquito species identification (Virginio et al. 2021). Another automated species identification project is Abuzz (Mukundarajan et al. 2017), which collects crowd-sourced data for vector tracking. Abuzz maintains a database with recordings of mosquito wing beat frequencies, which are used for identification (Mukundarajan et al. 2017).

While molecular data are included in many of the reviewed databases, there are platforms that curate genetic information beyond the scope of typical molecular surveillance initiatives. The Anopheles 1000 Genomes project (https://www.malariagen.net/mosquito/ag1000g) produces whole genome sequence datasets ("Ag1000G" 2022). Started in 2014, Ag1000G aims to use whole-genome deep sequencing on large numbers of wild-caught *Anopheles gambiae* (Giles, 1902) (Diptera: Culicidae) to improve understanding of natural genetic variation as it relates to ecology and malaria epidemiology. The Barcode of Life Data System (BOLD) (https://www.boldsystems.org) is a storage and analysis platform for DNA barcode records (Ratnasingham and Hebert 2007). Developed at the Centre for Biodiversity Genomics in Canada, the platform offers tools for management, analysis, and identification, in addition to assembly and organization of sequence data. Though not limited to arthropod vectors, BOLD provides an extensive resource for georeferenced molecular data.

## Information Flow and Overlaps

There are varying degrees of connectivity and interoperability between the databases outlined in this review (Fig. 1), reflecting diverse pathways and purposes for data collection, digitization, and sharing. Promoting interoperability, or the degree to which databases can be used together beyond an individual system, is important for ensuring that data are usable across platforms. Data standards, such as the Darwin Core metadata format, help ensure interoperability. Perhaps unsurprisingly, increased database connectivity can lead to overlaps and redundancies in records across platforms as information is shared across different databases or consolidated by large aggregators such as GBIF and SCAN. To illustrate information flow and overlap between databases and data streams, we searched for an individual occurrence record, tied to a museum specimen with a unique catalog number. *Aedes aegypti* (YPM ENT 999015), is a specimen collected on Vaca Key, Florida in 2003 that was deposited into the Yale Peabody Museum (YPM) Entomology Division collection (Fig. 2).

Running searches on the unique specimen number across the fully accessible databases in this review, we found that this record appears in GBIF, iDigBio, SCAN, and data associated with TPT. Although this is a clear instance of duplication across databases, it is also important to note that not all specimens are duplicated. For example, another record associated with a specimen deposited in the YPM entomology collection, *Ae. aegypti* (YPM ENT 683730) collected in Marathon, Florida in 1997, was retrieved from GBIF and iDigBio, but not SCAN. The degree to which databases overlap is beyond the scope of this work, and yet, is an important consideration for researchers planning to use occurrence records from multiple open data repositories.

Specimens deposited into museum collections, which are cataloged with unique identifiers as individuals or lots, represent an ideal scenario for tracing the flow of data. Such duplications may not be so easily identifiable from other sources (e.g., aggregated surveillance data, data contributed directly from projects, etc). Formatting and cleaning steps may also vary between databases, resulting in different data headings or dropped fields that may further complicate removal of duplicates. In the absence of unique identifiers, care should be taken when choosing criteria for duplicate removal, as seemingly unique attributes may have different formats across databases. For example, the number of provided decimal places may be different for the same geospatial locations (e.g., GPS points) across different databases, and thus when spatial de-duplication scripts are run, the same point may be interpreted as two different points instead. Problems with duplication may be more or less severe depending on modeling objectives. For example, presence-only species distribution modeling may be less vulnerable to duplication

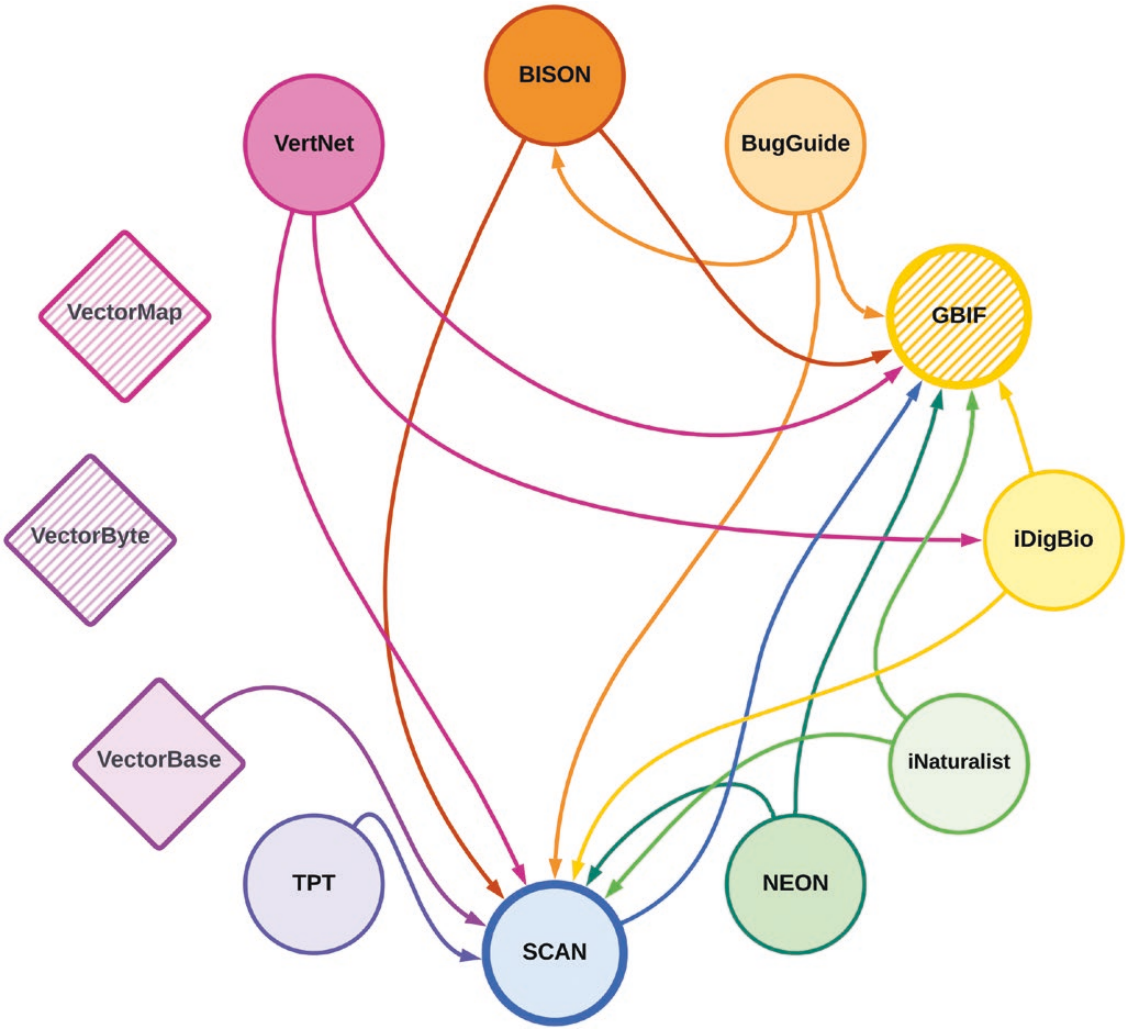


**Fig. 1.** Conceptual diagram showing connections between fully accessible databases of vector occurrence, highlighting the many pathways for data digitization and sharing. Vector databases are shown as diamonds, and databases that are inclusive of other taxa shown as circles; databases that do not currently export data to other platforms are shown with hashed fill, and bold outlines indicate major data aggregators. Note that connections indicate the availability of data products, and not necessarily direct data transfer events between platforms.


**Fig. 2.** *Aedes aegypti* specimen (YPM ENT 999015) collected from Vaca Key, FL, and deposited into the Yale Peabody Museum (photograph by Lawrence Gall, Yale University).

errors, as under common practices, occurrences undergo spatial duplication removal and thinning before modeling (Aiello-Lammens et al. 2015, Hijmans and Elith 2021). In contrast, population or forecast modeling may be particularly susceptible to duplication errors, which may be presumed to be abundance. Ultimately, these issues can be mitigated in the future by promoting the standardized capture of complete metadata across repositories.

In conclusion, efforts to increase the scope and accessibility of arthropod vector data over the past two decades have resulted in an ecosystem of online repositories that facilitate research on vector-borne disease systems (Table 1). However, due to diverse approaches and intents, the scope of the types of database architectures and contents can be confusing to navigate, and as capacity and support for the multitude of data efforts waxes and wanes, so too will accessibility and utility. The increasing availability of freely accessible data promotes the improvement and development of quantitative studies on arthropod disease vectors, and by extension, potential vector-borne disease risk. Nevertheless, easily accessible and interoperable data are not without potential caveats, such as duplication across repositories. These are potential issues that may be dually addressed by establishing and following best practices for data use, and by promoting and supplying sufficiently detailed metadata to accompany downloaded products. Indeed, an argument in favor of redundancy can be made, as this can help ensure sustainability of

**Table 1.** Databases and associated website links featured in this review, listed in order of appearance in text

| Database | Acronym | Website |
|---|---|---|
| Global Biological Information Facility | GBIF | https://www.gbif.org |
| Biodiversity Serving Our Nation | BISON | https://bison.usgs.gov/ipt/ |
| National Ecological Observatory Network | NEON | https://www.neonscience.org |
| Integrated Digitized Biocollections | iDigBio | https://www.idigbio.org |
| VertNet | – | http://vertnet.org |
| iNaturalist | – | https://inaturalist.org |
| BugGuide | – | https://bugguide.net |
| Symbiota Collections of Arthropods Network | SCAN | https://scan-bugs.org/ |
| Terrestrial Parasite Tracker | TPT | https://www.globalbioticinteractions.org/parasitetracker |
| VectorBase | – | https://vectorbase.org |
| VectorBaseMapVEu | – | https://vectorbase.org/popbio-map/web/ |
| VectorByte | – | https://www.vectorbyte.org |
| VectorMap | – | https://vectormap.si.edu |
| CDC ArboNet | – | https://wwwn.cdc.gov/arbonet/ |
| Ecological Database of the World's Insect Pathogens | EDWIP | https://edwip.ecology.uga.edu/<br>https://github.com/viralemergence/insectDisease |
| IR Mapper | – | https://www.irmapper.com |
| Malaria Atlas Project | MAP | https://malariaatlas.org |
| Malaria Threat Map | – | https://apps.who.int/malaria/maps/threats |
| Vectorborne Disease Surveillance System | VectorSurv | https://vectorsurv.org |
| VectorSurv Maps | – | https://maps.vectorsurv.org |
| VectorSurv Gateway | – | https://gateway.vectorsurv.org |
| European Network for Medical and Veterinary Entomology | VectorNet | https://vectornet.ecdc.europa.eu |
| Iowa Mosquito Surveillance | – | https://mosquito.ent.iastate.edu |
| VectorMap-GR | – | https://vectormap-gr.com |
| MosquitoDB | – | https://mosquitodb.io/mdb/login.php |
| ClinEpiDB | – | https://clinepidb.org |
| Mosquito Alert | – | http://www.mosquitoalert.com/en/ |
| Global Learning and Observations to Benefit the Environment | GLOBE | https://observer.globe.gov |
| Tick Report | – | https://www.tickreport.com/stats |
| WingBank | – | https://wingbank.butantan.gov.br |
| Anopheles 1000 Genomes Project | Ag1000G | https://www.malariagen.net/mosquito/ag1000g |
| Barcode of Life Data System | BOLD | https://www.boldsystems.org |

data products. When initiatives have time-limited funding streams, distributing holdings across databases and folding into new projects ensures their prolonged availability. In this piece, we provided a snapshot of the current vector data ecosystem, with a brief overview of aspects such as accessibility, scope, and data types.

## Acknowledgments

We were supported by CIBR: VectorByte: A Global Informatics Platform for studying the Ecology of Vector-Borne Diseases (SJR and CAL NSF DBI 2016265, and SSCR NSF-DBI-2016282) SJR was additionally supported by funding to Verena (viralemergence.org), including NSF BII 2021909 and NSF BII 2213854. We thank Neil Cobb for his helpful advice on the interrelatedness of databases. We thank Lawrence Gall for providing the specimen image.

## Author Contributions

CAL, SSCR and SJR conceived of this study, CAL, SSCR, and SJR drafted the paper, and CAL, SSCR and SJR reviewed and edited the final version.

## Competing Interests

The authors of this paper declare they are funded by the VectorByte Project and have received funding previous through the VectorBite RCN project. Rund has received funding, receives funding, or is affiliated with VectorMap, VectorBase, VEuPathDB, ClinEpiDB, and the TPT. These projects are described in this manuscript.

## Data Availability

While this paper describes data availability and access, there is no direct use of data in the manuscript.

## References Cited

Ag1000G. **2022**. Ag1000G. (https://www.malariagen.net/mosquito/ag1000g).

**Aiello-Lammens, M. E., R. A. Boria, A. Radosavljevic, B. Vilela, and R. P. Anderson. 2015**. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*. 38: 541–545.

**Amos, B., C. Aurrecoechea, M. Barba, A. Barreto, E. Y. Basenko, W. Bażant, R. Belnap, A. S. Blevins, U. Böhme, J. Brestelli, *et al.* 2022**. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.* 50: D898–D911.

**Barker, C. M., V. L. Kramer, and W. K. Reisen. 2010**. Decision support system for mosquito and arbovirus control in California. Earthzine. (https://earthzine.org/decision-support-system-for-mosquito-and-arbovirus-control-in-california/).

**Braks, M., F. Schaffner, J. M. Medlock, E. Berriatua, T. Balenghien, A. D. Mihalca, G. Hendrickx, C. Marsboom, W. Van Bortel, R. C. Smallegange, *et al.* 2022**. VectorNet: putting vectors on the map. *Front. Public Health*. 10: 809763.

**Braxton, S. M., D. W. Onstad, D. E. Dockter, R. Giordano, R. Larsson, and R. A. Humber. 2003**. Description and analysis of two internet-based databases of insect pathogens: EDWIP and VIDIL. *J. Invertebr. Pathol.* 83: 185–195.

**de Carvalho Gomes, H., G. D. Ali, J. Advait Francombe, A. Gkousis, R. Gloinson, E. Gunashekar, S. Leach, and S. Brandi Parkinson. 2021**. Digital technologies for infectious disease surveillance, prevention and control - a scoping review of the research literature 2015–2019. Stockholm: European Centre for Disease Prevention and Control.

**Cator, L. J., L. R. Johnson, E. A. Mordecai, F. E. Moustaid, T. R. C. Smallwood, S. L. LaDeau, M. A. Johansson, P. J. Hudson, M. Boots, M. B. Thomas, *et al.* 2020**. The role of vector trait variation in vector-borne disease dynamics. *Front. Ecol. Evol.* 8: 189.

**Delacour-Estrella, S., F. Collantes, I. Ruiz-Arrondo, P. M. Alarcón-Elbal, J. A. Delgado, R. Eritja, F. Bartumeus, A. Oltra, J. R. B. Palmer, J. Lucientes, *et al.* 2014**. Primeracita de mosquito tigre, Aedesalbopictus (Diptera, Culicidae), para Andalucía y primeracorroboración de losdatos de la aplicaciónTigatrapp, pp. 93–96. *In Anales de Biología*. Servicio de Publicaciones de la Universidad de Murcia.

**Foley, D. H., R. C. Wilkerson, and L. M. Rueda. 2009**. Importance of the "What," "When," and "Where" of mosquito collection events. *J. Med. Entomol.* 46: 717–722.

**Foley, D. H., R. C. Wilkerson, I. Birney, S. Harrison, J. Christensen, and L. M. Rueda. 2010**. MosquitoMap and the Mal-area calculator: new web tools to relate mosquito species distribution with vector borne disease. *Int. J. Health Geogr.* 9: 11.

**Fotakis, E. A., M. Orfanos, T. Kouleris, P. Stamatelopoulos, Z. Tsiropoulos, A. Kampouraki, I. Kioulos, K. Mavridis, A. Chaskopoulou, G. Koliopoulos, *et al.* 2021**. VectorMap-GR: a local scale operational management tool for entomological monitoring, to support vector control activities in Greece and the Mediterranean Basin. *Curr. Res. Parasitol. Vector Borne Dis.* 1: 100053.

GBIF Secretariat. **2020**. Establishing an effective GBIF participant node: concepts and general considerations. (https://docs.gbif.org/effective-nodes-guidance/1.0/en/).

Gigabyte: Vectors of human disease series. **2022**. Gigabyte: vectors of human disease series. (https://gigabytejournal.com/articles/series/GIGABYTE_SERIES_0002).

**Giraldo-Calderón, G. I., O. S. Harb, S. A. Kelly, S. S. Rund, D. S. Roos, and M. A. McDowell. 2021**. VectorBase.org updates: bioinformatic resources for invertebrate vectors of human pathogens and related organisms. *Curr. Opin. Insect Sci.* 50: 100860.

**Golding, N., A. L. Wilson, C. L. Moyes, J. Cano, D. M. Pigott, R. Velayudhan, S. J. Brooker, D. L. Smith, S. I. Hay, and S. W. Lindsay. 2015**. Integrating vector control across diseases. *BMC Med.* 13: 249.

**Hay, S. I., and R. W. Snow. 2006**. The malaria Atlas Project: developing global maps of malaria risk. *PLoS Med.* 3: e473.

**Hijmans, R. J., and J. Elith. 2021**. *Species distribution models*. Spatial Data Science (rspatial.org). Available online at https://rspatial.org/sdm

Institute of Medicine (US) Forum on Microbial Threats. **2008**. *Vector-borne diseases: understanding the environmental, human health, and ecological connections, workshop summary.* National Academies Press, Washington, DC.

**Kampen, H., J. M. Medlock, A. G. C. Vaux, C. J. M. Koenraadt, A. J. H. van Vliet, F. Bartumeus, A. Oltra, C. A. Sousa, S. Chouin, and D. Werner. 2015**. Approaches to passive mosquito surveillance in the EU. *Parasit. Vectors.* 8: 9.

**Kitsou, C., and U. Pal. 2022**. Vaccines against vector-borne diseases. *Methods Mol. Biol.* 2411: 269–286.

**Kiware, S. S., T. L. Russell, Z. J. Mtema, A. D. Malishee, P. Chaki, D. Lwetoijera, J. Chanda, D. Chinula, S. Majambere, J. E. Gimnig, *et al.* 2016**. A generic schema and data collection forms applicable to diverse entomological studies of mosquitoes. *Source Code Biol. Med.* 11: 4.

**Koita, O. A., O. K. Doumbo, A. Ouattara, L. K. Tall, A. Konaré, M. Diakité, M. Diallo, I. Sagara, G. L. Masinde, S. N. Doumbo, *et al.* 2012**. False-negative rapid diagnostic tests for malaria and deletion of the histidine-rich repeat region of the hrp2 gene. *Am. J. Trop. Med. Hyg.* 86: 194–198.

**Low, R., R. Boger, P. Nelson, and M. Kimura. 2021**. GLOBE mosquito habitat mapper citizen science data 2017-2020. *GeoHealth*. 5: e2021GH000436.

**Moy, B., R. Harrigan, and H. Godwin. 2018**. West Nile virus as a case study. *J. Environ. Health*. 80: 24–31.

**Moyes, C. L., A. Wiebe, K. Gleave, A. Trett, P. A. Hancock, G. G. Padonou, M. S. Chouaïbou, A. Sovi, S. A. Abuelmaali, E. Ochomo, *et al.* 2019**. Analysis-ready datasets for insecticide resistance phenotype and genotype frequency in African malaria vectors. *Sci. Data*. 6: 121.

**Mukundarajan, H., F. J. H. Hol, E. A. Castillo, C. Newby, and M. Prakash. 2017**. Using mobile phones as acoustic sensors for high-throughput mosquito surveillance. *Elife*. 6: e27854.

National Research Council (US) Board on Biology. **2010**. *Bioinformatics: converting data to knowledge: workshop summary.* National Academies Press (US), Washington (DC).

**Poelen, J. H., J. D. Simons, and C. J. Mungall. 2014**. Global biotic interactions: an open infrastructure to share and analyze species-interaction datasets. *Ecol. Inform.* 24: 148–159.

**Poelen, J. H., K. C. Seltmann, M. Campbell, S. A. Orlofske, J. E. Light, E. M. Tucker, J. R. Demboski, T. McElrath, C. C. Grinter, R. Diaz-Bastin, *et al.* 2021**. Terrestrial Parasite Tracker indexed biotic interactions and review summary. doi:10.5281/zenodo.7194486

**Ratnasingham, S., and P. D. N. Hebert. 2007**. BOLD: the Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes.* 7: 355–364.

**Ruhamyankaka, E., B. P. Brunk, G. Dorsey, O. S. Harb, D. A. Helb, J. Judkins, J. C. Kissinger, B. Lindsay, D. S. Roos, E. J. San, *et al.* 2019**. ClinEpiDB: an open-access clinical epidemiology database resource encouraging online exploration of complex studies. *Gates Open Res.* 3: 1661.

**Rund, S. S. C., K. Braak, L. Cator, K. Copas, S. J. Emrich, G. I. Giraldo-Calderón, M. A. Johansson, N. Heydari, D. Hobern, S. A. Kelly, *et al.* 2019a**. MIReAD, a minimum information standard for reporting arthropod abundance data. *Sci. Data*. 6: 40.

**Rund, S. S. C., I. K. Moise, J. C. Beier, and M. E. Martinez. 2019b**. Rescuing troves of hidden ecological data to tackle emerging mosquito-borne diseases. *J. Am. Mosq. Control Assoc.* 35: 75–83.

**Scheibner, J., A. Jobin, and E. Vayena. 2021**. Ethical issues with using Internet of Things devices in citizen science research: a scoping review. *Front. Environ. Sci. Eng. China.* 9: 629649.

**Secunda, P. M. 2004**. A Mosquito in the ointment: adverse HIPAA implications for health-related remote sensing research and a "reasonable" solution. *J. Space Law.* 30: 251–276.

**Suarez, A. V., and N. D. Tsutsui. 2004**. The value of museum collections for research and society. *Bioscience*. 54: 66–74.

**Sucaet, Y., J. Van Hemert, B. Tucker, and L. Bartholomay. 2008**. A web-based relational database for monitoring and analyzing mosquito population dynamics. *J. Med. Entomol.* 45: 775–784.

**Trivellone, V., W. Wei, L. Filippin, and C. H. Dietrich. 2021**. Screening potential insect vectors in a museum biorepository reveals undiscovered diversity of plant pathogens in natural areas. *Ecol. Evol.* 11: 6493–6503.

USGS. **2013**. Biodiversity information serving our nation (BISON). USGS - U.S. Geological Survey. (https://bison.usgs.gov/index.jsp?scientificName=Poa&ITIS=itis#about).

**Virginio, F., V. Domingues, L. C. G. da Silva, L. Andrade, K. R. Braghetto, and L. Suesdek. 2021**. WingBank: a wing image database of mosquitoes. *Front. Ecol. Evol.* 9:660941.

Walter Reed Biosystematics Unit. **2021**. Best practices guide: reporting entomological surveillance results to VectorMap. (https://vectormap.si.edu/downloads/guides/BestPractices_Data_Mgmt_Reporting_Guide_WRBU_2021.pdf)

Walter Reed Biosystematics Unit. **2022**. Vector hazard reports. (https://www.wrbu.si.edu/resources/vector_hazard_reports).