Journal of Healthcare Informatics Research
https://doi.org/10.1007/s41666-019-00055-2

SISS-Geo: Leveraging Citizen Science to Monitor Wildlife Health Risks in Brazil

Marcia Chame · Helio J. C. Barbosa · Luiz M. R. Gadelha Jr. · Douglas A. Augusto · Eduardo Krempser · Livia Abdalla

Received: 22 March 2018 / Revised: 9 June 2019 / Accepted: 20 June 2019

Abstract The well-being of human and wildlife health involves many challenges, such as monitoring the movement of pathogens; expanding health surveillance; collecting data and extracting information to identify and predict risks; integrating specialists from different areas to handle data, species and distinct social and environmental contexts; and, the commitment to bringing relevant information to society. In Brazil, there is still the difficulty of building a system that is not impaired by its large territorial extension and its poorly integrated sectoral policies. The Brazilian Wildlife Health Information System, SISS-Geo[1], is a platform for collaborative monitoring that intends to overcome the challenges in wildlife health. It aims at the integration and participation of various segments of society, encompassing: the registration of animals occurrences by citizen scientists; the reliable diagnosis of pathogens from the laboratory and expert networks; and computational and mathematical challenges in analytical and predictive systems, model interpretation, data integration and visualization, and geographic information systems. It has been successfully applied to support decision-making on recent wildlife health events, such as a Yellow Fever epizootic.

Keywords wildlife health · citizen science · information system

Marcia Chame · Douglas A. Augusto · Eduardo Krempser · Livia Abdalla
Oswaldo Cruz Foundation, Biodiversity & Wildlife Health Institutional Platform (PIBSS/Fiocruz), Rio de Janeiro, Brazil
E-mail: {marcia.chame,daa,eduardo.krempser,abdalla.livia}@fiocruz.br

Helio J. C. Barbosa · Luiz M. R. Gadelha Jr.
National Laboratory for Scientific Computing, Petrópolis, Brazil
E-mail: {hcbm,lgadelha}@lncc.br

[1] SISS-Geo is the abbreviation of "*Sistema de Informação em Saúde Silvestre Georreferenciado*" (which translates to "Georeferenced Wildlife Health Information System") and can be accessed at `http://www.biodiversidade.ciss.fiocruz.br` or `http://sissgeo.lncc.br` (in Portuguese).

1 Introduction

Environmental change, including climate change and biodiversity loss, are determining factors for the emergence of diseases originating from wildlife [15] and can be the source of the selective forces of new genetic variations that allow the disruption of biological barriers by pathogens and the increase in the potential for spread of diseases to humans. Although not considered appropriately in health surveillance policies, the situation is relevant, since the majority (60.3%) of infectious diseases circulate between humans and animals (zoonoses), of which 71.8% are caused by pathogens originating from wildlife [24]. Not to mention the alarming data from a recent study [55], which shows that the number of pathogens infecting humans and animals is vast and, more worryingly, they are growing over time.

These emergences are widely associated with areas most affected by natural and anthropogenic impacts, also composing the range of parameters that make social inequalities even more severe and unfair, with substantial repercussions and costs to health and quality of life [5,45]. Over the past decade, several studies have shown that biodiversity can affect both the dilution and dispersion of pathogens, as well as modulate their transmission rate [25,57,40].

However, studies and actions in the last century, despite the expansion of epidemiological knowledge, responded to specific disease emergence events in the human population, with some mitigation attempts. Considering the low ability to reverse climate change and the environmental impacts determined by human population growth, and the rate of production and consumption of natural resources, it seems reasonable to expect that the emergence of these diseases cannot be held back. This scenario is paradoxical in megadiverse countries, such as Brazil. While species richness results in richness of parasites that are associated to them, and therefore a potential risk, it is this complexity of species and their relationships that protect and stabilize the dynamics of transmission, reducing the outbreaks of diseases, one of the essential ecosystem services [26,34,9,46,53,28,3]. In this scenario, more than seeking effective responses to crises, there is a reason to pursue actions that anticipate problems so that one can mitigate them where possible, and quickly respond to them when prevention or mitigation fail.

This approach has been strengthened with international programs, such as "*One world, one health*" from the WHO/OIE and the 2011-2020 Strategic Plan of the Convention on Biological Diversity (CBD) [11] and strategically in governmental programs of developed countries. These already dedicate considerable resources and efforts to tracking pathogens, whether to prevent pandemics, such as the recent occurrences with influenza and Ebola viruses, the development of new drugs or even biological warfare concerns. There are programs and systems of surveillance of zoonoses in wild animals that have been acting essentially for the identification of new and old diseases, especially those of economic and conservationist interest and in the approach of One Health [13]. Most of these programs are: structured and maintained by governmental services, with professional personnel, collection protocols and

standardized diagnostic capacity (e.g., US Wildlife Disease Surveillance and Emergency Response), implemented by groups with scientific or conservation interest in one or a few species (e.g., World Conservation Society Health Program) or are based on the participation of trained farmers and hunters who are the first to come into contact with slaughtered animals, such as in Europe [49]. Except for studies of scientific interest and conservation of species, the other characteristics are not applicable in Brazil. In Brazil, systematized strategies for monitoring and predicting occurrences of diseases resulting from biodiversity are incipient. They follow a notification model about diseases that already occurred in humans or in a few species, which is insufficient for preventive action [7]. Firstly, there is no government system in place to monitor wildlife health consistently. Secondly, in Brazil, hunting is prohibited by law throughout the territory, except in particular places where vulnerable and traditional populations have the right to subsistence hunting. It should also be emphasized that the act of collecting biological samples for diagnosis imposes risks to the health of the person and therefore requires specific training and personal protection equipment. However, this is not consistent with the reality at the national level as well as with the vulnerability and low level of education of the majority of the population that lives in the forest of natural and anthropized environments.

The relationships that link biodiversity to health are complex because they are often indirect, scattered in space and time, and dependent on many forces [40]. The problem is not restricted to identifying species and their geographical distribution. In the context of the emergence of zoonoses, there are various species of pathogens, vectors, and hosts that modulate evolutionarily each other, their populational dynamics and composition, which collectively also undergo and react to environmental changes [25].

Therefore a multi-dimensional challenge is faced:

1. Sensitizing decision-makers about the need to monitor the movement of pathogens in wildlife before they impact humans, expanding health surveillance actions.
2. Building a mechanism that is not limited by the territorial extension of Brazil, the poorly integrated sectoral policies, and by other outbreaks or emergencies that absorb all the health staff.
3. How to integrate multiple skills, since this mechanism should contain specialists to handle data, species, and distinct social and environmental contexts.
4. How to effectively obtain, store, and manage data properly.
5. Modeling the risks from data to identify and predict them, as well as to extract the relevant information to convey it to society ultimately.

The first challenge is arguably the hardest one because it is mostly non-technical and involves dealing with politics. The ongoing strategy to sensitize decision-makers stands on two continuous actions: (i) getting in touch with decision-makers and, backed up by scientific studies, educating them about the benefits in terms of health, sustainability, economics, and politics from

taking preventive and predictive measures; (ii) presenting to decision-makers regularly how the SISS-Geo platform has been helping in disease prevention moreover, how the monitoring can be made both effective and inexpensive thanks to the network of volunteers and machine-learning based workflows. How the remaining challenges were dealt with in designing SISS-Geo will be explained further in the following sections.

As evidenced, data collection, monitoring and extraction of knowledge and information about wildlife health and its relationship to human health arise as challenging tasks involving several areas of knowledge, characterized as interdisciplinary activities aimed at modeling a dynamic and complex system. It is also clear that major areas of computing are mostly applicable in the context presented, such as computer modeling, machine learning, and parallel programming. However, their application is not apparent given the need to integrate information in different ways, the complexity and dimensionality of the data to be manipulated and the sensitivity involved in the use and dissemination of these data [42].

In this article, the Information System on Wildlife Health (SISS-Geo) is presented, a joint effort between the Oswaldo Cruz Foundation (Fiocruz) and the National Laboratory for Scientific Computing (LNCC), as an essential step for moving forward on the challenges posed. Its conception aimed at the integration and participation of various segments of society and encompasses: the registration of primary data by any person interested; the application of the concept of citizen science; the reliable diagnosis of pathogens circulating in wildlife that may potentially impact humans with the participation of laboratory and expert networks; the computational and mathematical challenges that include analytical and predictive systems, data mining, intensive processes, parallel programming, system integration, data (unstructured and heterogeneous) and information, geographic information systems (GIS), machine learning, meta-heuristics, and data visualization.

SISS-Geo is mainly characterized by managing its data in a spatially referenced environment. It aims to:

- provide, quickly and efficiently, the flow of information between (i) the Information Center for Wildlife at Fiocruz and the national system of health surveillance, with special contribution to the Strategic Information Center on Health Surveillance (CIEVS, Ministry of Health); (ii) the participatory networks in wildlife health and laboratories; (iii) the general population that wants to participate in the process; and (iv) the different biodiversity monitoring centers, as the MCTI (Ministry of Science, Technology and Innovation), ICMBio (Chico Mendes Institute for Biodiversity), MAPA (Ministry of Agriculture, Livestock and Supply), Embrapa (Brazilian Agricultural Research Corporation), etc.
- create, from the data and georeferenced information, warning and forecasting models on human and wildlife diseases in order to act as a sentinel system for emerging and reemerging diseases as well as provide the results of spatial modeling to scientific community and decision makers.

– allow for adequate means to integrate the georeferenced system with spatial databases partners from governmental and non-governmental partners.
– adapt to the metadata standard of the National Spatial Data Infrastructure (INDE) (`http://www.inde.gov.br`), aiming to provide, efficiently and with full compatibility, data related to wildlife health to the scientific community and the general population.

2 Design and Implementation of SISS-Geo

SISS-Geo is built upon four high-level modules, as illustrated in Figure 1. The

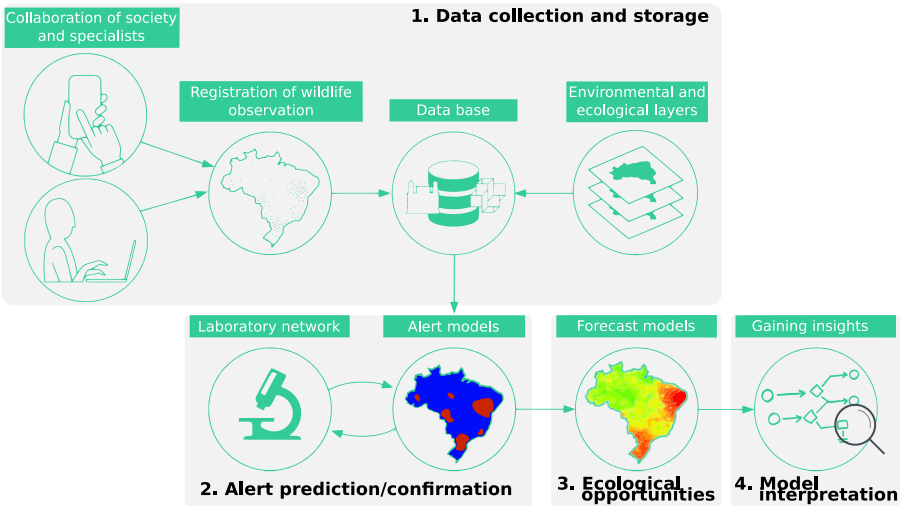

Fig. 1 Four modules of SISS-Geo, consisting of (1) data collection and storage, (2) alert prediction and confirmation, (3) forecast of ecological opportunities, and (4) model interpretation.

first one systematizes photographs and the capture of georeferenced field and observation records of animals, their physical conditions, and their surrounding environment, which are stored in a database (Sections 2.1and 2.2). Collaborators compile these observations through mobile applications, for Android (Figure 2), iOS, and in a Web interface (Figure 3). The second module analyzes the data to generate automated alert models that take into account territorial distances, time interval, similarity between taxonomic groups involved (notably for primates, Chiroptera, rodents, and carnivores, but not limited to them), the observed physical conditions of the animals in the field according to pre-categorized clinical patterns, and the environmental characteristics of the site where the animal was observed (Section 2.3.1). A georeferenced data explorer is available as well, allowing for multiple layers of information to be overlayed. Figure 4 illustrates a visualization where records (green), alerts (red), and biomes are overlayed in a map of Brazil.

Fig. 2 Screenshots of the SISS-Geo mobile application displaying the initial screen, main screen action buttons for taking photos and adding records, record description, and record map.

Fig. 3 Screenshots of the SISS-Geo Web application. Record details in the map (left), corresponding photo with a dead marmoset (right).

From the indication of importance and emergency generated by the alert model, the participatory and laboratory networks in wildlife and human health and environmental services established in the country are requested to collaborate on collecting biological samples from animals in the field and on providing reliable diagnoses. The reliable diagnosis feeds and validates the alert models which in turn, from the initial correlation of the environmental conditions of the occurrence, allows for the generation of forecast models of ecological opportunities for disease occurrence that may result from biodiversity loss, thus opening up a different research viewpoint. These actions comprise the third module (Section 2.3.2).

Finally, the fourth module approaches the challenge of understanding the relationships that govern the phenomenon in question, from the trained mod-

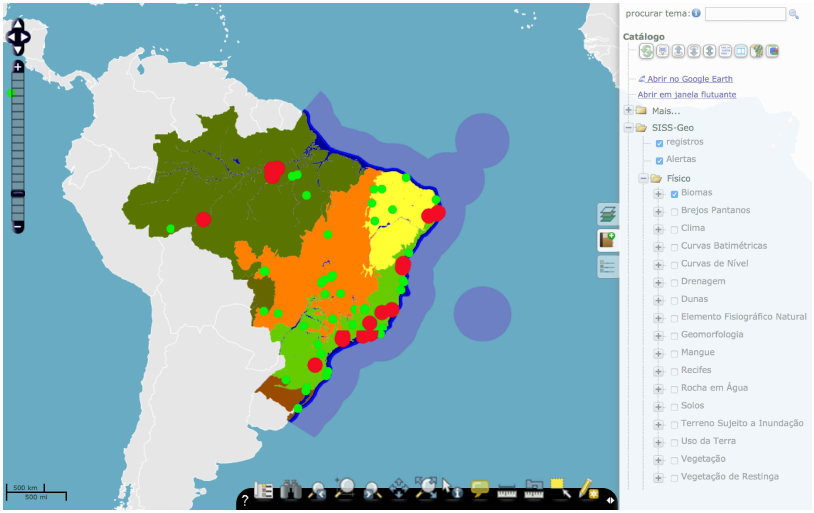

Fig. 4 Screenshots of the SISS-Geo georeferenced data explorer with options for displaying records, environmental, and socioeconomic layers in the right panel.

els. In this context, the model interpretation serves as the main hypotheses mechanism for further investigation and validation by experts (Section 2.3.3). The main components found in SISS-Geo can be categorized into four classes: wildlife health data management, GIS, machine learning, and wildlife health, in the next section.

It should be clear by now that in designing a platform whose (i) the primary source of data comes from citizens, i.e., it does not necessarily rely on the typically overburden health staff moreover, it is also not affected by sectoral policies at different administrative levels; also, (ii) has many components automated by smart workflows and machine learning, the platform is capable of covering the whole territorial extension of Brazil. Therefore, it overcomes the second challenge referred to in Section 1. In respect to the third challenge, SISS-Geo has been designed since the beginning to accommodate and integrate multiple use cases according to the role of each collaborator class, such as citizen scientists, specialists, laboratories and decision-makers (Fig. 1 and 5). The strength of SISS-Geo comes from the collaboration among these different users, with some providing the data (citizen scientists and specialists), others validating/processing it (specialists and laboratories) moreover, finally, a third group conveying the processed information to the academy (specialists) and society (decision-makers).

The components that had their implementations concluded correspond to the functionality that allows for gathering occurrence data from volunteers through the mobile or the Web application, storing the occurrences in the database, allowing specialists and laboratories to manipulate the occurrences, and allowing for the data to be geographically explored. These components are fully functional and are deployed as mobile applications (for Android and

iOS), Web applications (for manipulating occurrences and for geographic exploration), and a database. They correspond to the wildlife health data management and GIS classes and are described in sections 2.1 and 2.2. In section 2.3, a methodology is presented for generating alerts using machine learning techniques; this functionality is still under implementation.

2.1 Data management in wildlife health

To monitor changes in biodiversity, one needs to collect, document, store, and analyze indicators of the spatial and temporal distribution of species, as well as information on how they interact with each other and with the environment they live in [30]. The development and implementation of mechanisms to produce these indicators [37] depend on access to reliable data from field surveys, automated sensors, biological collections, and from the academic literature. This data is usually available in various institutions that use different formats and identifiers, which makes it a challenging data integration task. The methods and techniques used to manage and analyze this data define a research area often called Biodiversity Informatics [21,39]. Some initiatives for establishing metadata and data publishing standards, such as EML [16] and Darwin Core [56], were able to present standard vocabularies used to describe concepts of biodiversity. Although these vocabularies cover only a fraction of the possible concepts, they allow institutions to publish their data about biodiversity using the same format, and for their automatic collection and processing by aggregator systems.

Through the use of these standards, SISS-Geo can collect species occurrence data provided by various contributors, as well as providing data stored in its database to the community at large in an easy to use format. Darwin Core has been extended to include concepts on specific topics, such as information about interactions and pollinators (*Darwin Core Extension for Interactions*) and on species profiles (*Plinian Core*) [36]. It would be essential to evaluate and propose an extension of the standard to include information about wildlife health on species observation records, which is typically carried out in the context of the Biodiversity Information Standards[2] (TDWG) organization.

SISS-Geo is a biodiversity informatics platform and, as such, it allows for users to upload species occurrence records. In SISS-Geo, these records are enriched with additional attributes, provided by the user, to describe the health condition of the respective individuals. The term *occurrence* is used in this work to refer to the observation of an individual that apparently carries a disease, which is a particular case of a species occurrence as commonly defined in the biodiversity informatics literature. Its geographical scope is limited to Brazil, and the users are given by citizen scientists and specialists. A relational database was conceptually modeled and implemented for SISS-Geo comprising occurrences of organisms along with associated information about their health

[2] `http://www.tdwg.org`

condition. Standard operations for creating, reading, updating and deleting information are enabled by mobile and Web applications that allow for both citizen scientists and system managers to interact with the system (Figure 5 describes SISS-Geo use cases).

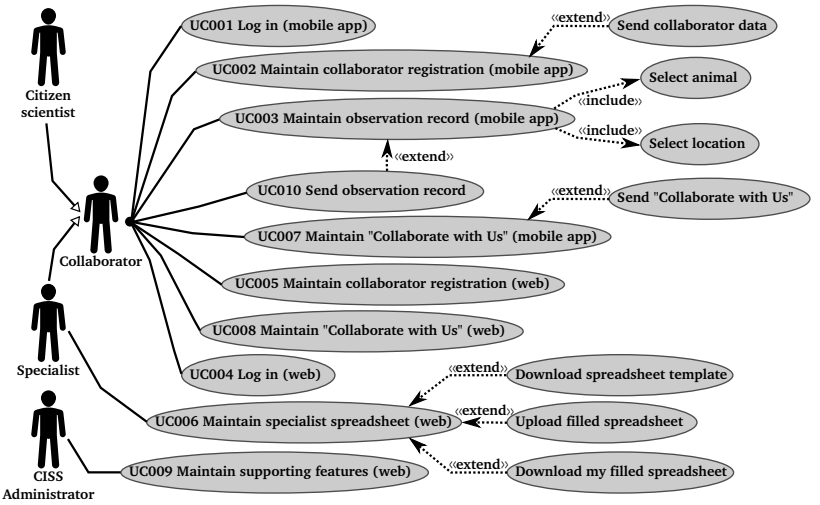

Fig. 5 Use cases of SISS-Geo displaying various possible interactions between users and functions of SISS-Geo.

As can be observed in its database schema in Figure 6, SISS-Geo stores information about wildlife health occurrences (Occurrence). These occurrences usually have an animal (Animal), a collaborator (Collaborator) and a location (Location) associated with them. Specialists can require samples (Sample) related to the occurrence to be collected, which are going to be analyzed (Analysis) in the laboratory (Laboratory) network. Data stored in this database is consumed by mathematical models that can produce and confirm wildlife health alerts (Alert).

The architecture of SISS-Geo is described in Figure 7. It is comprised of the following components: a mobile application, a Web application server, a database server, and high-performance computing (HPC) resources. As described in the use cases diagram in Figure 5, citizen scientists use the mobile application to request, for instance, the upload of their observations or queries to be executed. These requests are forwarded to the Web application server which connects to the database server to answer these requests. Administrative users and specialists can access the Web application server directly also to send requests to SISS-Geo. Finally, the Web application server can invoke the execution of computationally-intensive analyzes on high-performance computing resources. A complete list of use cases is described in Figure 5.

The approach used to tackle the fourth challenge mentioned in Section 1, of effectively obtaining, storing, and managing data is based on following the

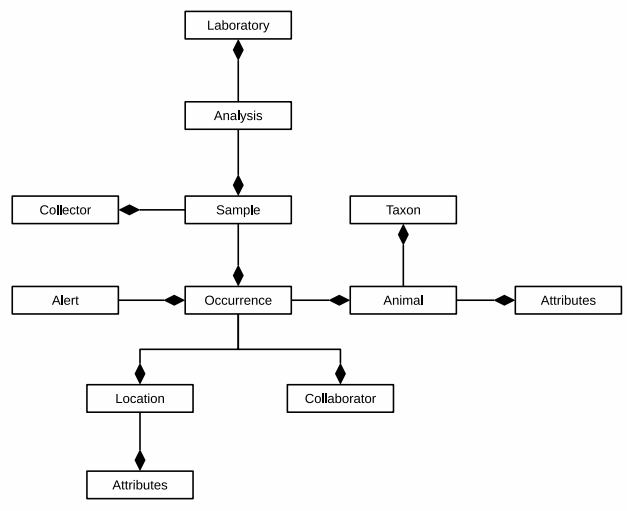

Fig. 6 Overview of the various entities and relationships that comprise the database schema of SISS-Geo.

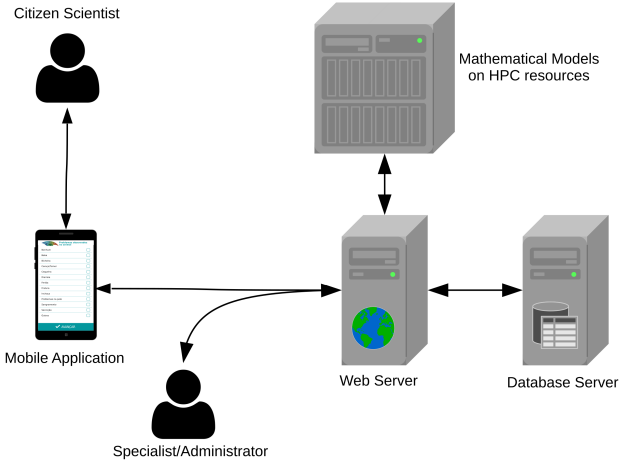

Fig. 7 Architectural view of SISS-Geo displaying components of the system and their interactions with citizen scientists, specialists, and system administrators.

best practices for scientific data management, especially from the biodiversity informatics community. The conceptual model of SISS-Geo's database follows established standards, such as Darwin Core [56], and the Ecological Metadata Language (EML) [16]. Following the example of citizen science initiatives, such as eBird [52], SISS-Geo can obtain massive valuable data from volunteers that use its mobile application in Android and iOS platforms. As described in the next subsection, this data is combined with other datasets, and it is used in the alert prediction model proposed in subsection 2.3.

2.2 Geoprocessing

Spatial and geographical visualization are fundamental conditions for the management of information today. It is often difficult due to the need for normalization, update, and access to qualified data. In studies of infectious diseases, the spatialization of data needs additionally to consider populational pulses and fluctuations determined by several factors such as seasonality, reproductive periods, migrations, among others [35].

SISS-Geo aims to generate relevant and reliable information that can support decision processes of the Brazilian Ministries of Health, Environment, Agriculture, Livestock, and Supply providing subsidies for more agile and timely decision making.

Because it is an innovative project, the functionality developed is not straightforward, and it was often not available in similar initiatives. The construction of new methodologies and the use of different types of geographic technologies that can meet the expectations and objectives of SISS-Geo is therefore necessary. The GIS Infrastructure (GI) of SISS-Geo has strategic importance in this process, in which there is a need to overcome challenges related to quality control of spatial data, modeling spatialization based on machine learning and the dissemination of models in the form of dynamic maps on the Internet.

The data-driven modeling of diseases occurrence based on socio-environmental variables in SISS-Geo uses a broad diversity of spatial data, such as land use and vegetation cover (Mapbiomas collection 2.3[3]); temperature and precipitation (Global Precipitation Mission - GPM[4] and Worldclim[5]); geomorphology, soil types, climatic zones, degree of urbanization, highways, mineral exploration areas, biomes, and conservation units (Brazilian Institute of Geography and Statistics - IBGE[6]); demographic density (NASA's Socioeconomic Data and Applications Center[7]); altimetry (NASA's Aster GDEM[8]). Since these data come from different sources (Brazilian and other national and global sources), they have different scales, reference systems, and mapping methodologies. Therefore, they were pre-processed and structured for integration into a geographic database. It is used both to consume information/data and to store the modeling results in the form of geographically distributed models. The data used as input for modeling are obtained from the overlapping of wildlife occurrence records and environmental, social, and human impact databases. Depending on the location of the records, spatial relationships of the types *intersect*, *within*, *close*, *crosses*, and the like can be established.

All pre-processing tasks, performed on over one hundred gigabytes of data, were carried out in QGIS [41]. At this stage, it was necessary to standardize

[3] http://mapbiomas.org

[4] https://pmm.nasa.gov/data-access/downloads/gpm

[5] http://www.worldclim.org

[6] https://www.ibge.gov.br

[7] http://sedac.ciesin.columbia.edu

[8] https://asterweb.jpl.nasa.gov/gdem.asp

the cartographic characteristics of geographic data, correct topological errors, clean duplication of information, and standardize the structure of the attribute table. In general, the data were divided into two groups: vector data and raster data. All data in raster format was converted to vector format in order to be compatible with the internal software package which expects this format as input.

Knowing that part of the thematic data used was produced in small and medium scale (1:1,000,000, 1:500,000, 1:250,000), which provide a limited level of detail and accuracy, the verification methodology of spatial relations adopted areas of influence (buffers) on the occurrence points of the animal species. It brought flexibility for spatial queries, allowing to identify the context of socio-environmental features on which the animal was observed.

Other spatial and temporal information is requested to the user and added to the database as observation site ("local scale") attributes to enhance species observation records used for data-driven modeling.

The geoprocessing infrastructure also needs to make available the results, alerts, and prediction models produced by SISS-Geo to the public domain according to the Brazilian Information Access Act, except for sensitive information. Therefore, adequating the geographic information system for the Web environment, which provides SISS-Geo's results in the form of dynamic/interactive maps and graphical statistics[9], is an ongoing development. An advantage of this technology is the ease of handling, analysis, and interpretation of models by the end user, as well as operating system independence and interaction with desktop systems and other Internet systems (interoperability).

2.3 Machine Learning

SISS-Geo embraces machine learning techniques to fulfill the fifth challenge mentioned in Section 1, leading to risk mapping and the understanding of factors related to the emergence of diseases. These products are vital for the genuine purpose of SISS-Geo because they account for the main avenue of conveying information to decision-makers and society. The first component (Section 2.3.1) deals with real-time alert prediction, which intends to target health authorities for further verification and diagnostic of alerts. The second and third components (Sections 2.3.2 and 2.3.3) aim, respectively, at building models and extracting knowledge from them in order to advance the understanding of associations between socio-environmental factors and suitability for disease occurrence, which are of vital importance for specialists, decision-makers, and society.

2.3.1 Grouping of observation records and alert prediction

When a wild animal is observed, its physical condition and surrounding environment are recorded in SISS-Geo, either by experts or volunteers. These

[9] `http://morcego.siss.lncc.br/i3geo/interface/black_ol.htm`

records are grouped with other related records (previously reported) resulting in a collection of events characterizing a phenomenon. This is the grouping stage and, although it may sound trivial, it involves the challenge of conceiving/training models with the discriminative capacity to recognize similarities and dissimilarities between events, based on criteria such as spatial and temporal distance between records, the similarity between species and the reported physical conditions, among others. This flow of learning is summarized in Figure 8.

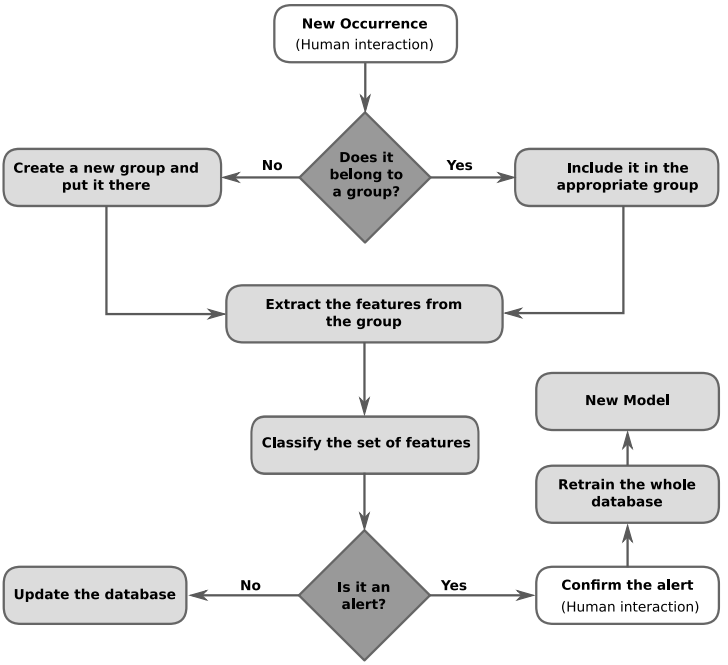

Fig. 8 Machine learning flow of SISS-Geo, starting with a new occurrence, going through decision and analysis steps, and finishing in either a database update to record the occurrence or in an alert confirmation followed by the creation of a new model.

The second part consists of modeling the characteristics of observation records that make them more or less relevant, i.e., training the alert model. It means predicting the severity of records according to information brought by events and the geographic/environmental context. For example, a record involving an animal in isolation exhibiting symptoms is less severe, in general, than occurrences containing similar events but covering groups of animals. Of course, in real situations, the characterization of an alert situation is usually much less noticeable, commonly taking into consideration many factors for decision making. In some cases, a single record is sufficient to generate an alert, such as the registration of a wild canid with symptoms of rabies and non-human primates with Yellow Fever symptoms.

It can be seen that the activities mentioned above refer to the grouping and data classification task, typical of machine learning, and well known for the wide variety of approaches and methodologies. They are therefore complex tasks, both by nature as well as by the large volume of data expected for the system[10].

However, the challenges of grouping and classification that are present in SISS-Geo go beyond the classic challenges of these tasks.

Phenomenon characterization. The characterization of what defines a group of events (phenomenon) lies in the problem of non-conventional similarity measurement formulation (e.g., not necessarily Euclidean). Grouping rules based on expert experience are a reasonable alternative but has as shortcoming the limited formalization of knowledge and, consequently, the potential for the introduction of unwanted biases. Another approach is to treat this problem as a machine learning process, aiming at the training of similarity models: given a new record and the existing ones, determine to which group it belongs — or whether it characterizes a new group. The process is characterized as supervised learning, since it is possible to determine reliably, *a priori* or *a posteriori*, which records belong to which phenomenon, either by empirical tests or by expert confidence.

Feature extraction. Once constituted the phenomena, it is necessary to evaluate them as to the potential threat to wildlife health and its possible outbreak in humans, as phenomena alone do not necessarily constitute alert situations. In this sense, information characterizing a group of events needs to be extracted and provided to the alert prediction model. The difficulty is thus to derive statistics which better represent the phenomenon described by the group in order to maximize the performance of the prediction model; in other words, raise the necessary information to facilitate the learning process. Experts recommend the use of certain statistics, such as the type and quantity of affected animals, number and frequency of occurrences, among others; however, the space of possible features goes well beyond that and could be used to improve predictive performance. Thus, an open question is how to exploit this vast space automatically? An interesting line of research and a potential solution to this challenge is the investigation of automatic feature extraction methods [20,19]. In a nutshell, the task can be cast as a supervised machine-learning problem by taking as independent variables the union of the information of all events in a group and, as the corresponding dependent variable, whether or not an alert was issued at the time—of course, this requires the existence of pre-labeled alerts. Then, a machine-learning algorithm can be applied to learn a function (or set of functions) that maximizes the correlation between groups' content and alert prediction; this optimized function can be understood as the *extracted feature*.

Alert prediction model. Although its use in the system is similar to sufficiently known methods described in the literature, the alert prediction model

[10] After all, it is an ambitious system that aims to aggregate and store records on wildlife health of a vast country.

is probably the most strategic component of SISS-Geo's intelligence. The viability of the system is fundamentally based on the accuracy of the prediction model, both in detecting *true positives* (alerts) as *true negatives* (non-alerts). The failure to detect an alert condition (false negative) can result in severe consequences to wildlife, environmental, and human health. On the other hand, false positives would overwhelm the relatively small network of laboratories and experts responsible for confirming or denying alerts (more details below). In this sense, methods that combine multiple models (*ensemble methods*) usually produce more accurate and robust solutions. Therefore they are promising candidates as training algorithms for prediction models [44]. Still, since the large portion of the system's data has no associated class, that is, phenomena whose alert predictions have not yet been confirmed, the semi-supervised learning is an interesting approach due to its ability also to leverage unlabelled instances in the training process [8].

Alert confirmation. Another key component of SISS-Geo—on which all others depend—is the process of alert confirmation. This step is the second (and last) time a human interacts in the process, the other being the upload of the observation record. As expected, a great challenge and bottleneck result from the need for direct human participation in the confirmation procedure, either in the field or laboratory; it is an expensive and slow process, even considering the extensive network of qualified collaborations linked to SISS-Geo. When there are more alerts issued by the prediction model than the capacity of experts and the laboratory network to confirm them, the phenomena need to be prioritized. In this situation, one can think of prioritizing the phenomena associated with alerts (1) by *alert severity* weighted by the *confidence* of prediction; or (2) by *relevance to regions of great interest*, be it social, environmental or economic. However, a strategy focused on the medium and long term is the prioritization of confirmation (or denial) of alerts with greater potential for improving the accuracy of the prediction model. This line of research is recent, and it is called *active learning* [50]. The same method can also be used in possible cases of false negatives, thus avoiding the possibility of degeneration of the prediction model[11]: the phenomena predicted as non-alerts but that are promising from the learning point of view would be subject to confirmation (of the non-alert condition) by an expert.

2.3.2 Prediction of Ecological Opportunities for Disease Occurrence

Another line of fundamental importance in SISS-Geo is the prediction of scenarios and environments that favor ecological opportunities for disease occurrence arising from wildlife or, put differently, raising scenarios conducive to the occurrence of a particular event, such as an outbreak of a disease.

In short, trained alert models can be used to evaluate different scenarios and characterize those potentially susceptible. From these, environmental,

[11] Consider the extreme situation where all the predictions are *non-alerts*, including both true as false negatives. Since, in principle, only the cases of alerts are of interest and subject to confirmation, in this scenario, the model would be doomed to degeneration.

social, and human and animal health variables are taken (see Section 2.2), leading to a set of instances that share the status of "abnormality", according to the alert model predictions. Then, data-driven models are built over this set in order to estimate a distribution of socio-environmental variables related to alerts. Finally, the resulting models can be applied for *predictive* or *descriptive* purposes. While in the former the goal is to assign a degree of suitability for disease occurrence in the geographic space (regions), the latter aims at the understanding of the factors associated with the disease occurrence, akin to the discussion in Section 2.3.3.

In order to construct these predictive/descriptive models, methods for linking of the mentioned variables, such as the ones applied to ecological niche modeling [51] or, preferably, the less specific traditional machine learning methods can be applied in this context. Since the alert models are trained based on confirmed/denied events (Figure 8), what the disease occurrence models will be reconstructing is not only the (realized) niche of the observed animals but hopefully the environmental and climatic parameters that favor the realized niche of the pathogen, which potentially include portions of the niche of its components including vectors and hosts (since non-human primates coexist with other species) [38]. Take, for instance, the Sylvatic Yellow Fever disease. When an alert is issued (due perhaps to an observed high number of non-human primate deaths), specialists will confirm or deny the alert. In this case, it is the same as confirming (or not) the circulation of the YF virus among non-human primates, which in turn is connected with the circulation of YF mosquitoes.

It is worth observing that this kind of modeling outputs the *suitability* for disease occurrence, not the actual *probability* of occurrence. In other words, the model measures how close a given region's socio-environmental variables are to the distribution of the corresponding variables of regions that had confirmed alerts [38].

2.3.3 Gaining insights through model interpretation

An essential feature of symbolic modeling methods, such as decision trees, rule extraction algorithms and meta-heuristic genetic programming [27], is that they reveal in human-readable form the existing relationships between the input and output data.

The potential of this class of models to aid experts is remarkable in the analysis and understanding of the phenomenon investigated, leading to a man-machine interaction: the model suggests hypotheses that best fit the data while the expert validates them.

In order to gain meaningful insights from the model, it is necessary to accurately define its structure/language or, in other words, to incorporate expert knowledge properly. While doing that, care should be taken to find the ideal balance between *bias*, usually resulting from structural simplicity of the model, and *variance*, an issue usually associated with structurally more complex models.

3 Evaluation

As of February 2018, SISS-Geo was downloaded more than a thousand times from the Google Play store and had an average rating of 4.8 out of 5 stars. Even though the potential number of observations related to wildlife health usually being a fraction of the population of a species, SISS-Geo has 3,014 records in its database performed by 1,881 citizen scientists. Its Web interface has been accessed 4,463 times. These records correspond to 764 mammals, 815 birds, 383 reptiles, 227 amphibians, 47 fish, and 540 not identified. Table 1 lists the ten most recorded taxonomic groups in SISS-Geo. It is important to emphasize that the records were uploaded by volunteer collaborators that often do not have taxonomic knowledge, which can have adverse effects on data quality. To tackle this issue and improve wild animal monitoring, which can lead to better assertive models for the emergence of zoonoses, SISS-Geo has developed a tool for expert-supported record validation. Figure 9 shows the geographic distribution of the observations recorded by SISS-Geo that are georeferenced.

Table 1 Ten most recorded taxonomic group in SISS-Geo - data until March 2018

Common name	Number of records	%
Not identified	540	18.3
Birds (Other)	417	14.1
Birds (Penelopes, Seriemas, Toucans)	112	8.6
Amphibians	242	8.2
Snakes	189	6.4
Marmosets and Tamarins	177	6.0
Turtles	101	5.9
Lizards	75	3.4
Birds of prey	72	2.5
Capybaras	64	2.2

SISS-Geo integrates data-based computational modeling, development, and high-performance computing. It was selected in 2014 as the best project [6] in the "Health" category of the *Grand Challenges of Computing* event of the Brazilian Computer Society. In 2017, SISS-Geo received the National Biodiversity Prize from the Brazilian Ministry of the Environment[12]. It allows the monitoring of wildlife and can support the identification of zoonoses, such as the Yellow Fever outbreaks, which in its sylvatic cycle circulates among non-human primates. The fact that monkeys become ill or die before there are human cases of Yellow Fever causes the surveillance of outbreaks, such as the recent one [12,33], in these animals to be of vital importance in the control and prevention of the disease. The collaboration of the population is critical because prevention actions can be improved and streamlined, and everyone

[12] `http://www.mma.gov.br/component/k2/item/10443-pr&` (In Portuguese)

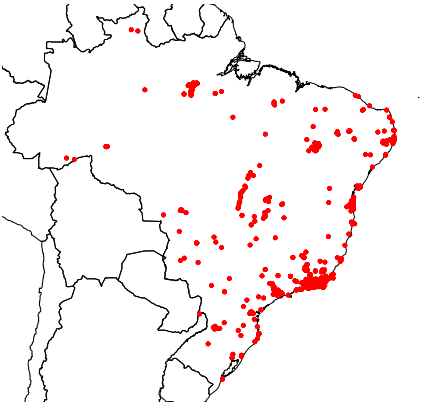

Fig. 9 Geographic distribution of records (red dots) in Brazil until March 2018.

will benefit. With the participation of ordinary people, the application makes available, in real time, the occurrences of dead or diseased animals for public health and biodiversity conservation, assisting the Epizootics Surveillance System in Nonhuman Primates (PNH), of the Brazilian Ministry of Health, and records of dead monkeys are reported to the responsible bodies investigating the cases. The information recorded in SISS-Geo serves to generate computational models for predicting zoonoses and for the adoption of preventive measures. Tables 2 and 3 list the recorded conditions and the most recorded abnormalities in SISS-Geo, respectively.

Table 2 Recorded conditions in SISS-Geo until March 2018

Condition	Number of records	%
Normal behavior	2,168	71.2
Dead animal	697	22.9
Strange behavior	112	3.7
Sick animal	67	2.2

Some of the observations performed with SISS-Geo triggered alerts and contributed to biodiversity conservation actions, such as: (i) 59 dead turtles were recorded in the south of the Brazilian state of Bahia in November 2017, generating a notification to the responsible environmental agency and a legal notice to those involved in predatory fishing in the area; (ii) observations of dead foxes with rabies in the Northeast were able to support decision-making by health surveillance agencies; (iii) 73 dead monkeys were recorded **in 2016**

Table 3 Recorded abnormalities in SISS-Geo until March 2018

Abnormality	Number of records	%
None	2,377	81.3
Wound	103	3.5
Other	101	3.5
Fracture	75	2.6
Blindness	74	2.5
Bleed	58	2.0
Skin problems	36	1.2
Swell	32	1.1
Myiasis	25	0.9
Secretion	18	0.6
Drool	16	0.5
Lump or Tumor	6	0.2
Diarrhea	3	0.1

during the recent Yellow Fever epizooty, which directed health surveillance actions in the field.

The outbreak of yellow fever, which occurred in 2016 and spread throughout southeastern Brazil [54, 33], was evidenced by SISS-Geo from the recording of non-human primates in Minas Gerais and other states. Among the various prevention and control actions, the Health Surveillance Secretariat of the Ministry of Health of Brazil carried out five training courses with all the agents and stakeholders involved in the surveillance of Yellow Fever in all the states of the country. In these training sessions, SISS-Geo was presented and offered to agents and managers as a monitoring tool for zoonoses epizootics [1]. Other training activities were carried out with community health agents, park guards and civil defense agents directly involved in the actions of human vaccination and collection of biological samples of non-human primates to confirm cases, in addition to the capture of animals. The use of SISS-Geo, although unofficial so far, since it is necessary to restructure the national flow of information, has been adopted as an additional tool in surveillance, especially for the ability to generate georeferencing, photographs, and real-time information. As a result of this work, working groups, municipal agents, and collaborators from 25 Brazilian states record animals, which has already helped to inform about 200 deaths of non-human primates throughout the country.

SISS-Geo also contributed to the monitoring of species on the IUCN Red List of Threatened Species, with the availability of the location and information of some species already registered as *Panthera onca*, *Puma concolor*, *Tapirus terrestris*, *Myrmecophaga tridactyla*, *Bradypus torquatus*, *Chrysocyon brachyurus*, *Chelonia mydas*, *Leontopithecus chrysomelas*, *Alouatta guariba guariba*, and *Crax blumenbachii*.

As seen, the SISS-Geo platform, including its data and methodologies, allows analyses that can go beyond the initial planning: from the monitoring of specific groups of animals to its complete adaptation to new contexts. Thus,

examples of the expansion of SISS-Geo to new scenarios can be seen by the use of data already collected, its tools, or even all its computational framework.

In this sense, projects that rely on the records of the platform to support biodiversity monitoring, such as in Serra dos Órgãos National Park (PARNASO - *Parque Nacional da Serra dos Órgãos* in Rio de Janeiro State), are already in progress. Besides that, the collected records can also be used in many other scenarios, for instance: estimating species distribution along with their health status and training models of automatic species identification from SISS-Geo's images.

Besides, the SISS-Geo computing framework, designed to integrate locality information, photographic and animal records, is easily replicated, taking advantage of all the tools and methodologies developed and applied in its design. An example of this reuse is the project under development as a partnership between Fiocruz, the National Center for Flora Conservation of the Rio de Janeiro Botanical Garden Research Institute (CNCFlora/JBRJ) and the Rio de Janeiro State Secretary of the Environment (SEA-RJ). It aims to adapt SISS-Geo to a citizen-scientist platform for searching for rare plants, within the context of the project "*Campanha Procura-se*[13]". By simply adapting the information from the module "Animal" to a "Plant" one, it was possible to replicate most of the previously described concepts and flows.

4 Related Work

He et al. [22] present the eMammal framework for wildlife monitoring supported by citizen scientists. Animal images collected with camera traps are sent to its database where visual animal recognition techniques are applied. The species identification recommendations generated are reviewed by citizen scientists and, subsequently, by experts. The resulting validated records are made available to wildlife and ecological researchers. eBird [52] also leverages the capability of citizen scientists to gather bird observation records. Automated data quality filters are used to support species identifications performed by citizen scientists. iNaturalist [23] is another biodiversity citizen-science initiative available as both a mobile and Web application. Volunteers can record and identify species observations that can be validated by other users and biologists. After an observation is validated, it is annotated as "research grade" and uploaded to GBIF. The World Organization for Animal Health (OIE), which maintains the World Animal Health Information Database (WAHIS) Interface[14], is accessible online and contains updated information on disease outbreaks. However, most of OIE's pertinent and relevant information relates to infectious agents that have an impact on livestock production and human health, when the two situations are interlinked. As an example, there is no notification of Yellow Fever epizootics in humans and non-human primates, and this information was only requested to Brazil in 2017 after thousands of deaths

[13] `http://dspace.jbrj.gov.br/jspui/handle/doc/95`

[14] `http://www.oie.int/wahis_2/public/wahid.php/Wahidhome/Home`

of people and non-human primates. In WAHIS, space for "other diseases" of irrelevant paper to the economy was added only in recent years. It is important to note that the information provided by the OIE comes from member countries in half-yearly reports. Brazil has the National Animal Health Information System (SIZ)[15], however, the database refers to the mandatory notification diseases described in Normative Instruction No. 50, 23rd of September, 2013. The national list does not include pathogens that have no interest for animal production, although there is also a field where any notification can be made. Therefore, in Brazil, there is no system for collecting and systematizing wildlife diseases, one of the main reasons for the development of SISS-Geo, created by a project in partnership with government agencies of livestock production and the environment.

More general biodiversity databases exist at the global, national, and ecosystem levels. GBIF [14] gathers species observation data on a global scale. In February 2018, it had 54 national nodes. Along with other types of participants, GBIF gathers data from 1,152 institutions, totaling approximately a billion records. SiBBr [18] is the Brazilian GBIF node, publishing species occurrence records and providing an ecological niche modeling portal [47]. BaMBa [29] is a biodiversity database that focuses on marine ecosystems that is also integrated with GBIF. These systems use the IPT tool [43] to extract observation records from local databases, export them to Darwin Core [56], and publish them on GBIF.

SISS-Geo is both a citizen science application and a biodiversity database. eBird, eMammal, and iNaturalist while being citizen science applications as well, do not provide tools for data analysis as SISS-Geo plans to do with the application of machine learning techniques to generate wildlife health alerts following the methodology proposed in subsection 2.3. GBIF, SiBBr, and BaMBa focus on data mobilization and publication and do not directly provide tools for enabling the participation of citizen scientists.

5 Conclusion

The proposal was inspired by the desire to make public and seek reinforcements for a long walk that brings together researchers, experts from multiple areas and society so that, through computing, information and disease prevention actions reach the most remote regions of the country. It emerges from many years of practice of field research in the Brazilian semi-arid region, where relevant information on diseases in wild animals have been lost or dispersed, and the lack of systematization turned necessary actions impossible both for the containment of diseases in humans, as for conservation of species.

SISS-Geo was born out of efforts to create innovative and integrated actions for the mainstreaming of biodiversity in the sectors of the country. It

[15] http://www.agricultura.gov.br/assuntos/sanidade-animal-e-vegetal/animal-animal/epidemiologia/ingles/animal-health-information-system

integrates the actions of the Oswaldo Cruz Foundation (Fiocruz) in "Public-Private Actions for Biodiversity Project" – PROBIO II[16], coordinated by the Brazilian Ministry of the Environment, and developed by FUNBIO, Embrapa, the Brazilian Ministries of Agriculture and Livestock, Health, and Science Technology and Innovation, the Botanical Garden of Rio de Janeiro, ICMBio and Fiocruz. The National Laboratory for Scientific Computing joined the Fiocruz project and ensured its execution in a long-term knowledge-building partnership.

By automating the search for occurrence patterns, the information reaches more efficiently citizens nationwide, from the general population through experts, as well as provides the opportunity for the acquisition of knowledge about the possible patterns and parameters that contribute to the occurrence of diseases. In the medium- and long-term, it also builds the capacity of researchers to develop complex modeling in the ecology of diseases that can exploit geographic information in order to improve accuracy. Moreover, occurrence patterns yield data that can assist national policy on health and biodiversity conservation.

It should be pointed out that all the described design decisions and techniques embodied in SISS-Geo could also be readily adapted to other similar settings. Besides the increasingly popular concept of leveraging citizen scientists to propel a collaborative monitoring tool, which would fit countless different scenarios provided they can either tolerate some inaccuracy or have a mechanism to validate user input, the proposed machine-learning flow is sufficiently generic to cover a wide range of related contexts. For instance, it is rather common the case in which a phenomenon cannot be recorded all at once, but incrementally in time and space, possibly by different collaborators, as in the jigsaw puzzle; here, everything discussed in Section 2.3.1 would be potentially useful. Another potential contribution to other settings is the discussion about feature extraction from groups followed by the alert prediction/confirmation, which could play a central role in situations—especially in those related to unattended monitoring—where certain events should trigger alerts.

In the context of SISS-Geo, the incorporation provenance information is planned to allow the alert generation process to be traceable, meaning that one can recover the data, configuration parameters, people and computational activities involved. This enables many applications, such as assessing the quality of the alerts generated, verifying compliance with governmental regulations, and the reproducibility [10,4] of the alert generation process. Provenance information [32,31], which contains details about the planning and execution of computational processes, such as scientific workflows, describing the processes and data involved in the generation of its results may be used to facilitate this task. They allow an accurate description of how a computational process was planned, which is called *prospective provenance*, and what occurred during execution, which is called *retrospective provenance*. Some applications

[16] `http://www.funbio.org.br/probioii`

of provenance include reproducibility of computational processes for validation, sharing and reuse of knowledge, data quality evaluation, and attribution of scientific results. One of the concepts commonly captured in provenance is *causality*, which is given by the existing dependency relationships between computational activities and data sets. These dependencies can derive, by transitivity, dependencies between data sets and between processes.

An application programming interface (API) is being implemented and will serve the data stored in SISS-Geo to other systems. The installation of an instance of IPT [43] is also planned and will allow this same data to be exposed in the Darwin Core [56] standard along with EML [16] metadata . This will enable global and national biodiversity databases, such as GBIF and SiBBr [18], to collect the species occurrence records stored in SISS-Geo. As a result of the need to integrate information on wildlife, human and livestock health, several conversations have been made with the OIE's Brazilian focal point in the government's Ministry for Agriculture, Livestock and Food Supply of Brazil. The plan is for information from SISS-Geo to feed the SIZ (Brazilian National Animal Health Information System) database, which will subsequently power the WAHIS database. Since 2017 the transfer of information has been done informally, with the reporting of notifications of dead and sick animals. Systematic integration in the various national information systems, both human health and livestock production, also depends on many political and legal advances and, in particular, on the strengthening of the structure that supports the laboratory diagnosis of pathogens that do not appear in the notifiable diseases nor for humans, or livestock. The integration data from additional data providers that are relevant to the application area of SISS-Geo is planned, following the rules and national legislation.

In one way or another, the authors believe that the SISS-Geo platform addresses the five challenges mentioned in Section 1:

1. Decision-makers can be sensitized to the importance of wildlife monitoring through multiple avenues, such as (i) scientific communication (as this document itself) and (ii) models of diseases occurrence that are capable of anticipating outbreaks. SISS-Geo is being used in ongoing collaborative work with the Secretariat of Health Surveillance (Brazilian Ministry of Health) to generate data-driven Yellow Fever models.
2. Regarding the second challenge, being an easy-to-use GIS-based tool that leverages citizen scientists as the primary source of data collection from wildlife health, SISS-Geo is not impaired by large territorial extension neither is it dependent on sectoral policies and government health staff. The independent network of experts and laboratories takes care of alert confirmation; concomitantly, the active-learning approach proposed in Section 2.3.1 would hopefully minimize the human resource involved in this task.
3. SISS-Geo has been designed since the beginning as a platform aimed at integrating seamlessly multiple agents and skills. Each agent, whether it is a citizen scientist, specialist, laboratory or decision-maker, has a well-

defined role in SISS-Geo. Citizen scientists and specialists can both upload records of observations to SISS-Geo; the difference is that records provided by experts tend to be more comprehensive and reliable. After that, they are validated and, if alerts are predicted, the network of experts and laboratories are asked to confirm or deny them. From the confirmed/denied data, models that correlate factors and occurrence are built via machine-learning techniques with the aid of experts. Then, occurrence cases, alerts, and models are communicated to decision-makers, enabling them to make informed decisions on possibly imminent outbreaks.

4. For obtaining, storing, and managing data properly, this is effectively accomplished in SISS-Geo thanks to its architecture with a dedicated Web server, database server, and HPC resources (Figure 7). There are ongoing efforts into making the SISS-Geo mobile application capable of operating flawlessly in offline mode, which is very common in remote areas of the country.
5. Finally, the fifth challenge concerning the identification and prediction of risks from data, as well as the extraction and communication of relevant information, are also part of the SISS-Geo workflow, tackled respectively by the tasks of modeling of disease occurrence (Section 2.3.2) and model interpretation (Section 2.3.3).

Availability

Web interface:
`http://sissgeo.lncc.br` and `http://www.biodiversidade.ciss.fiocruz.br`
Geographical explorer:
`http://morcego.siss.lncc.br/i3geo/interface/black_ol.htm`
Mobile application (Android):
`https://play.google.com/store/apps/details?id=siss.ui`
Mobile application (iOS):
`https://itunes.apple.com/br/app/siss-geo/id1291912325`

Acknowledgments

This work is funded by the Global Environment Facility (GEF) among World Bank, Caixa Econômica Federal, Brazilian Biodiversity Fund (Funbio) and Fiocruz for the development of the "National Biodiversity Mainstreaming and Institutional Consolidation Project" coordinated by the Brazilian Ministry of the Environment.

This is a post-peer-review, pre-copyedit version of an article published in the Journal of Healthcare Informatics Research. The final authenticated version is available online at: `http://dx.doi.org/10.1007/s41666-019-00055-2`.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

1. Alves, R.: Surto de Febre Amarela: A experiência da resposta no Brasil (2017). URL https://www.arca.fiocruz.br/bitstream/icict/22895/7/14h30-RenatoAlves.pdf
2. Augusto, D.A., Barbosa, H.J.: Accelerated parallel genetic programming tree evaluation with OpenCL. Journal of Parallel and Distributed Computing **73**(1), 86–100 (2013). DOI 10.1016/j.jpdc.2012.01.012. URL https://doi.org/10.1016/j.jpdc.2012.01.012
3. Betts, A., Gray, C., Zelek, M., MacLean, R. C., and King, K. C.: High parasite diversity accelerates host adaptation and diversification. Science, 360(6391), 907911 (2018). DOI 10.1126/science.aam9974. URL https://doi.org/10.1126/science.aam9974
4. Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B.D., Nabrzyski, J., Stodden, V., Taylor, I.J., Turk, M.J., Turner, K.: Computing Environments for Reproducibility: Capturing the Whole Tale. Future Generation Computer Systems (2018). DOI 10.1016/j.future.2017.12.029. URL https://doi.org/10.1016/j.future.2017.12.029
5. CBD: Subsidiary Body on Scientific Technical and Technological Advice. Consideration of Issues in Progress: Health and Biodiversity. Convention on Biological Diversity. Sixteenth meeting. Tech. rep. (2014).
6. Chame, M., Barbosa, H.J.C., Gadelha, L., Augusto, D.A., Krempser, E., Abdalla, L.: Sistema de Informação em Saúde Silvestre - SISS-Geo. In: III Seminário Grandes Desafios da Computação no Brasil - Fase 2. SBC (2014).
7. Chame, M., Labarthe, N.: Saúde Silvestre e Humana: Experiências e Perspectivas. Fundação Oswaldo Cruz (Fiocruz) (2013).
8. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised learning. MIT Press (2010).
9. Civitello, David J. and Cohen, Jeremy and Fatima, Hiba and Halstead, Neal T. and Liriano, Josue and McMahon, Taegan A. and Ortega, C. Nicole and Sauer, Erin Louise and Sehgal, Tanya and Young, Suzanne and Rohr, Jason R.: Biodiversity inhibits parasites: Broad evidence for the dilution effect. Proceedings of the National Academy of Sciences (2015). DOI 10.1073/pnas.1506279112.
10. Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Bras, Y.L., Lemoine, F., Mareuil, F., Ménager, H., Pradal, C., Blanchet, C.: Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Generation Computer Systems **75**, 284–298 (2017). DOI 10.1016/j.future.2017.01.012. URL https://doi.org/10.1016/j.future.2017.01.012.
11. Convention on Biological Diversity (CBD): Decision Adopted by the Conference of the Parties to the Convention on Biological Diversity at its Tenth Meeting. X/2. The Strategic Plan for Biodiversity 2011–2020 and the Aichi Biodiversity Targets. UNEP/CBD/COP/DEX/X/2 (2010). URL https://www.cbd.int/doc/decisions/cop-10/cop-10-dec-02-en.pdf.
12. Couto-Lima, D., Madec, Y., Bersot, M.I., Campos, S.S., Motta, M.d.A., Santos, F.B.d., Vazeille, M., Vasconcelos, P.F.d.C., Lourenço-de Oliveira, R., Failloux, A.B.: Potential risk of re-emergence of urban transmission of Yellow Fever virus in Brazil facilitated by competent Aedes populations. Scientific Reports **7**(1), 4848 (2017). DOI 10.1038/s41598-017-05186-3. URL https://doi.org/10.1016/10.1038/s41598-017-05186-3.
13. Cunningham, A., Daszak, P., Wood, J. One Health, emerging infectious diseases and wildlife: two decades of progress? Philosophical Transactions of the Royal Society B: Biological Sciences **372**: 20160167 (2017). DOI 10.1098/rstb.2016.0167. URL https://doi.org/10.1098/rstb.2016.0167.
14. Edwards, J.L.: Research and Societal Benefits of the Global Biodiversity Information Facility. BioScience **54**(6), 486 (2004). DOI 10.1641/0006-3568(2004)054[0486:RASBOT]2.0.CO;2. URL http://bioscience.oxfordjournals.org/content/54/6/485.full.

15. Estrada-Peña, A., Ostfeld, R.S., Peterson, A.T., Poulin, R., de la Fuente, J.: Effects of environmental change on zoonotic disease risk: an ecological primer. Trends in parasitology **30**(4), 205–14 (2014). DOI 10.1016/j.pt.2014.02.003. URL https://doi.org/10.1016/j.pt.2014.02.003.
16. Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M.: Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. Bulletin of the Ecological Society of America **86**(3), 158–168 (2005). DOI 10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2. URL http://www.esajournals.org/doi/abs/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.
17. Fodor, I.: A survey of dimension reduction techniques. Tech. rep., Center for Applied Scientific Computing, Lawrence Livermore National Laboratory (2002).
18. Gadelha, L., Guimarães, P., Moura, A.M., Drucker, D.P., Dalcin, E., Gall, G., Tavares, J., Palazzi, D., Poltosi, M., Porto, F., Moura, F., Leo, W.V.: SiBBr: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira. In: VIII Brazilian e-Science Workshop (BRESCI 2014). Proc. XXXIV Congress of the Brazilian Computer Society (2014).
19. Guo, L., Rivero, D., Dorado, J., Munteanu, C.R., Pazos, A.: Automatic feature extraction using genetic programming: An application to epileptic EEG classification. Expert Systems with Applications **38**(8), 10,425–10,436 (2011). DOI 10.1016/j.eswa.2011.02.118. URL https://doi.org/10.1016/j.eswa.2011.02.118.
20. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature extraction: foundations and applications. Springer-Verlag (2006).
21. Hardisty, A., Roberts, D., Addink, W., Aelterman, B., Agosti, D., Amaral-Zettler, L., Ariño, A.H., Arvanitidis, C., Backeljau, T., Bailly, N., Belbin, L., Berendsohn, W., Bertrand, N., Caithness, N., Campbell, D., Cochrane, G., Conruyt, N., Culham, A., Damgaard, C., Davies, N., Fady, B., Faulwetter, S., Feest, A., Field, D., Garnier, E., Geser, G., Gilbert, J., Grosche, Grosser, D., Herbinet, B., Hobern, D., Jones, A., de Jong, Y., King, D., Knapp, S., Koivula, H., Los, W., Meyer, C., Morris, R.A., Morrison, N., Morse, D., Obst, M., Pafilis, E., Page, L.M., Page, R., Pape, T., Parr, C., Paton, A., Patterson, D., Paymal, E., Penev, L., Pollet, M., Pyle, R., von Raab-Straube, E., Robert, V., Robertson, T., Rovellotti, O., Saarenmaa, H., Schalk, P., Schaminee, J., Schofield, P., Sier, A., Sierra, S., Smith, V., van Spronsen, E., Thornton-Wood, S., van Tienderen, P., van Tol, J., O'Tuama, E., Uetz, P., Vaas, L., Vignes Lebbe, R., Vision, T., Vu, D., De Wever, A., White, R., Willis, K., Young, F.: A decadal view of biodiversity informatics: challenges and priorities. BMC ecology **13**(1), 16 (2013). DOI 10.1186/1472-6785-13-16. URL https://doi.org/10.1186/1472-6785-13-16.
22. He, Z., Kays, R., Zhang, Z., Ning, G., Huang, C., Han, T.X., Millspaugh, J., Forrester, T., McShea, W.: Visual Informatics Tools for Supporting Large-Scale Collaborative Wildlife Monitoring with Citizen Scientists. IEEE Circuits and Systems Magazine **16**(1), 73–86 (2016). DOI 10.1109/MCAS.2015.2510200. URL https://doi.org/10.1109/MCAS.2015.2510200.
23. Heberling, J. M., Isaac, B. L.: iNaturalist as a tool to expand the research value of museum specimens. Applications in Plant Sciences **6**(11): e01193 (2018). DOI 10.1002/aps3.1193. URL https://doi.org/10.1002/aps3.1193.
24. Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P.: Global trends in emerging infectious diseases. Nature **451**(7181), 990–993 (2008). DOI 10.1038/nature06536. URL https://doi.org/10.1038/nature06536.
25. Keesing, F., Holt, R.D., Ostfeld, R.S.: Effects of species diversity on disease risk. Ecology Letters **9**(4), 485–498 (2006). DOI 10.1111/j.1461-0248.2006.00885.x. URL https://doi.org/10.1111/j.1461-0248.2006.00885.x.
26. Keesing, Felicia and Ostfeld, Richard S.: Is biodiversity good for your health?. Science **349**(6245), 237–236 (2015). DOI 10.1126/science.aac7892.
27. Koza, J.R., R., J.: Genetic programming : on the programming of computers by means of natural selection. MIT Press (1992). URL https://dl.acm.org/citation.cfm?id=138936.
28. Lafferty, K. D., Allesina, S., Arim, M., Briggs, C. J., De Leo, G., Dobson, A. P., Dunne J. A., Johnson, P. T. J., Kuris, A. M., Marcogliese, D. J., Martinez, N. D., Memmott,

J, Marquet, P. A., McLaughlin, J. P., Mordecai, E. A., Pascual, M., Poulin, Robert, Thieltges, D. W.: Parasites in food webs: the ultimate missing links.. Ecology Letters, **11**(6), 533-546 (2008). DOI 10.1111/j.1461-0248.2008.01174.x. URL https://doi.org/10.1111/j.1461-0248.2008.01174.x.

29. Meirelles, P.M., Gadelha Jr., L.M.R., Francini-Filho, R.B., Leão, R.d.M., Amado-Filho, G.M., Bastos, A.C., Paranhos, R.P.d.R., Rezende, C.E., Swings, J., Siegle, E., Neto, N.E.A., Leitão, S.N., Coutinho, R., Mattoso, M., Salomon, P.S., Valle, R.A.B., Pereira, R.C., Kruger, R.H., Thompson, C., Thompson, F.L.: BaMBa: towards the integrated management of Brazilian marine environmental data. Database **2015** (2015). DOI 10.1093/database/bav088. URL https://doi.org/10.1093/database/bav088.
30. Michener, W.K., Jones, M.B.: Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution **27**(2), 85–93 (2012). DOI 10.1016/j.tree.2011.11.016. URL https://doi.org/10.1016/j.tree.2011.11.016.
31. Mondelli, M. L., Magalhes, T., Loss, G., Wilde, M., Foster, I., Mattoso, M., Katz, D., Barbosa, H., de Vasconcelos, A. T. R., Ocaa, K., Gadelha, L. M. R.: BioWorkbench: a high-performance framework for managing and analyzing bioinformatics experiments. PeerJ **6**: e5551 (2018). DOI 10.7717/peerj.5551. URL https://doi.org/10.7717/peerj.5551.
32. Moreau, L., Groth, P.: Provenance: An Introduction to PROV, vol. 3. Morgan & Claypool (2013). DOI 10.2200/S00528ED1V01Y201308WBE007.
33. Moreira-Soto, A., Torres, M.C., Lima de Mendonça, M.C., Mares-Guia, M.A., Damasceno dos Santos Rodrigues, C., Fabri, A., Cardoso dos Santos, C., Machado Araújo, E.S., Fischer, C., Ribeiro Nogueira, R.M., Drosten, C., Sequeira, P.C., Drexler, J.F., Bispo de Filippis, A.M.: Evidence for Multiple Sylvatic Transmission Cycles During the 2016-2017 Yellow Fever Virus Outbreak, Brazil. Clinical Microbiology and Infection (2018). DOI 10.1016/J.CMI.2018.01.026. URL https://doi.org/10.1016/J.CMI.2018.01.026.
34. Ostfeld, Richard and Keesing, Felicia: Biodiversity and Disease Risk: the Case of Lyme Disease. Conservation Biology **14**, 722–728 (2000). DOI 10.1046/j.1523-1739.2000.99014.x.
35. Ostfeld, R.S., Glass, G., Keesing, F.: Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in Ecology & Evolution **20**(6), 328–336 (2005). DOI 10.1016/j.tree.2005.03.009. URL https://doi.org/10.1016/j.tree.2005.03.009.
36. Pando, F.: How species interactions are managed in Plinian Core: Status and questions. Proceedings of TDWG **1**:e20556 (2017). DOI 10.3897/tdwgproceedings.1.20556. URL https://doi.org/10.3897/tdwgproceedings.1.20556.
37. Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M.: Ecology. Essential biodiversity variables. Science (New York, N.Y.) **339**(6117), 277–8 (2013). DOI 10.1126/science.1229931. URL https://doi.org/10.1126/science.1229931.
38. Peterson, A. T.: Mapping Disease Transmission Risk: Enriching Models using Biogeography and Ecology. JHU Press (2014).
39. Peterson, A.T., Soberón, J., Krishtalka, L.: A global perspective on decadal challenges and priorities in biodiversity informatics. BMC ecology **15**(1), 15 (2015). DOI 10.1186/s12898-015-0046-8. URL https://doi.org/10.1186/s12898-015-0046-8.
40. Poulin, R.: Network analysis shining light on parasite ecology and diversity. Trends in Parasitology **26**(10), 492–498 (2010). DOI 10.1016/j.pt.2010.05.008. URL https://doi.org/10.1016/j.pt.2010.05.008.
41. QGIS Development Team: QGIS Geographic Information System. Open Source Geospatial Foundation (2009). URL http://qgis.osgeo.org.
42. Ryser-Degiorgis, M.: Wildlife health investigations: needs, challenges and recommendations. BMC Veterinary Research **9**:223 (2013). DOI 10.1186/1746-6148-9-223. URL https://doi.org/10.1186/1746-6148-9-223.
43. Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L., Desmet, P.: The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE **9**(8), e102,623

(2014). DOI 10.1371/journal.pone.0102623. URL https://doi.org/10.1371/journal.pone.0102623.
44. Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review **33**(1-2), 1–39 (2010). DOI 10.1007/s10462-009-9124-7. URL https://doi.org/10.1007/s10462-009-9124-7.
45. Romanelli, C., Cooper, D., Campbell-Lendrum, D., Maiero, M., Karesh, W.B., Hunter, D., Golden, C.D.: Connecting global priorities: biodiversity and human health: a state of knowledge review. World Health Organization and Secretariat of the Convention on Biological Diversity (2015).
46. Rottstock, Tanja and Joshi, Jasmin and Kummer, Volker and Fischer, Markus: Higher plant diversity promotes higher diversity of fungal pathogens, while it decreases pathogen infection per plant. Ecology **95**(7), 1907–1917 (2014). DOI 10.1890/13-2317.1.
47. Sánchez-Tapia, A., de Siqueira, M.F., Lima, R.O., Barros, F.S.M., Gall, G.M., Gadelha, L.M.R., da Silva, L.A.E., Osthoff, C.: Model-R: A Framework for Scalable and Reproducible Ecological Niche Modeling. In: High Performance Computing: 4th Latin American Conference, CARLA 2017. Communications in Computer and Information Science, vol. 796, pp. 218–232. Springer (2018). DOI 10.1007/978-3-319-73353-1_15. URL https://doi.org/10.1007/978-3-319-73353-1_15.
48. Santos, S.D.R., Huinca, S.C.M.: Considerações sobre a utilização da PEC Padrão de Exatidão Cartográfica nos dias atuais. In: III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias de Geoinformação. Recife (2009).
49. Schulz, K., Calba, C., Peyre,M., Staubach, C., Conraths, F. Hunters' acceptability of the surveillance system and alternative surveillance strategies for classical swine fever in wild boar - a participatory approach. BMC Veterinary Research (2016) 12:187. DOI 10.1186/s12917-016-0822-5. URL https://doi.org/10.1186/s12917-016-0822-5.
50. Settles, B.: Active Learning Literature Survey. Tech. rep., University of Wisconsin-Madison (2009)
51. Sillero, N.: What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecological Modelling **222**(8), 1343–1346 (2011). DOI 10.1016/j.ecolmodel.2011.01.018.
52. Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J.W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.A., Merrifield, M., Morris, W., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K.V., Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W.K., Wood, C.L., Yu, J., Kelling, S.: The eBird enterprise: An integrated approach to development and application of citizen science. Biological Conservation **169**, 31–40 (2014). DOI 10.1016/J.BIOCON.2013.11.003. URL https://doi.org/10.1016/J.BIOCON.2013.11.003.
53. Sutherst, R. W.: The vulnerability of animal and human health to parasites under de global change. International Journal for Parasitology **31**, 933–948 (2001). DOI 10.1016/S0020-7519(01)00203-X. URL https://doi.org/10.1016/S0020-7519(01)00203-X.
54. Vasconcelos, P., Monath, T.: Yellow Fever Remains a Potential Threat to Public Health. Vector-Borne and Zoonotic Diseases **16**(8) (2016). DOI 10.1089/vbz.2016.2031. URL http://doi.org/10.1089/vbz.2016.2031.
55. Wardeh, M., Risley, C., McIntyre, M.K., Setzkorn, C., Baylis, M.: Database of host-pathogen and related species interactions, and their global distribution. Scientific Data **2**, 150,049 (2015). DOI 10.1038/sdata.2015.49. URL https://doi.org/10.1038/sdata.2015.49.
56. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D.: Darwin Core: an evolving community-developed biodiversity data standard. PloS one **7**(1), e29,715 (2012). DOI 10.1371/journal.pone.0029715. URL https://doi.org/10.1371/journal.pone.0029715.
57. Xavier, S.C.d.C., Roque, A.L.R., Lima, V.d.S., Monteiro, K.J.L., Otaviano, J.C.R., Ferreira da Silva, L.F.C., Jansen, A.M.: Lower Richness of Small Wild Mammal Species and Chagas Disease Risk. PLoS Neglected Tropical Diseases **6**(5), e1647 (2012). DOI 10.1371/journal.pntd.0001647. URL https://doi.org/10.1371/journal.pntd.0001647.