# A Geographical Approach for Metadata Quality Improvement in Biological Observation Databases

Daniel Cintra Cugler
Institute of Computing
University of Campinas
13.083-970 – Campinas – SP – Brazil
danielcugler@ic.unicamp.br

Claudia Bauzer Medeiros
Institute of Computing
University of Campinas
13.083-970 – Campinas – SP – Brazil
cmbm@ic.unicamp.br

Shashi Shekhar
Department of Computer Science
University of Minnesota
55455 – Minneapolis – MN – USA
shekhar@cs.umn.edu

Luís Felipe Toledo
Fonoteca Neotropical & Museu de Zoologia
Institute of Biology
University of Campinas
13.083-862 - Campinas, SP, Brazil
toledolf2@yahoo.com

***Abstract*—This paper addresses the problem of improving the quality of metadata in biological observation databases, in particular those associated with observations of living beings, and which are often used as a starting point for biodiversity analyses. Poor quality metadata lead to incorrect scientific conclusions, and can mislead experts in their analyses. Thus, it is important to design and develop methods to detect and correct metadata quality problems. This is a challenging problem because of the variety of issues concerning such metadata, e.g., misnaming of species, location uncertainty and imprecision concerning where observations were recorded. Related work is limited because it does not adequately model such issues. We propose a geographic approach based on expert-led classification of place and/or range mismatch anomalies detected by our algorithms. Our work is tested using a case study with the Fonoteca Neotropical Jacques Vielliard, one of the 10 largest animal sound collections in the world.**

## I. Introduction

Our work concerns the curation of databases containing records of observations of living beings. An observation concerns the occurrence of an organism or set of organisms detected at a given place and time according to some methodology. In other words, "an observation represents an assertion that a particular entity was observed and that the corresponding set of measurements were recorded (as part of the observation)" [8]. Observation databases store a variety of data, at multiple spatial and temporal scales, including images, maps, sounds, texts and so on. In several domains, the reliability of metadata is a key concern for scientists because errors can lead to incorrect conclusions that may ripple across an entire study and beyond. For example, in biodiversity studies, metadata errors regarding a single species can affect understanding not just of the species, but of wider ecological interactions. Metadata quality improvement in such a scenario is challenging not only due to the intrinsic heterogeneity of such data, but also because of the many scientists who intervene in specifying and curating metadata, for distinct kinds of spatial and temporal granularities.

Some recent publications on curation of scientific metadata – e.g., [2] [29] – are mostly directed towards citizen provided information, which is known to be less reliable than data entered by domain experts. However, our experiments show that, no matter how much effort scientists put into curating data, there is still considerable margin for errors. This tends to grow with data volume. For instance, a simple set of checks performed by our group on another scientist-curated data set showed that roughly 20% of the records still contained errors, such as typos in species names, or lack of standardization.

Some errors in metadata are specific to the domain (e.g., misidentified species). Others are found in all kinds of metadata, and include problems such as duplicated records, typographical errors, data outside the correct range, incomplete data fields. Typically, errors in metadata are detected through various data cleaning and curating methods [9] [25]. The growing size of biological observation databases means that data cleaning and curating processes have become ever more arduous and time-consuming. Our work aims to develop new computational methods to ease this burden.

Metadata quality improvement in such databases is challenging because of the observation methodologies adopted. Such observations often result from many scientific expeditions undertaken along the years. As remarked by [8], for instance, since observation records depend on such teams, they suffer from both schema and semantic heterogeneity (i.e., structure and content). Thus, not only is there a large percentage of legacy records, but heterogeneity caused by methodological variations in observations. Related Computer Science work in data cleaning in this domain is limited, being mostly concerned with fixing typing and numeric errors, but without performing further correlations. Even in cases where filters are provided to take into account the location where species are expected to live (e.g., [29]), there is little concern with uncertain and imprecise descriptions of locations (e.g., via place names and region names), or with outdated species classification.

To address these limitations, we provide a novel perspective. We propose a geographic approach for metadata quality improvement in biological observation databases, as detailed

in Figure 1. In our case study with animal sound observations, for example, our approach enables detection of anomalies in both species' reported geographic distributions and in species' identification. Our goal is to support biologists in detecting metadata errors that are domain-related, and that need expert knowledge, thereby alleviating the burden of manual curation.

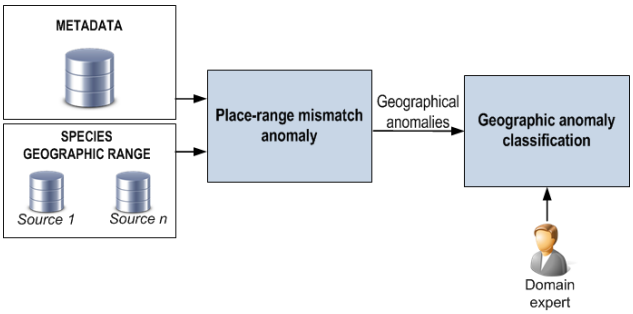


Fig. 1. Overview of our geographical approach.

Our approach is evaluated using a case study at the Fonoteca Neotropical Jacques Vielliard (FNJV) [14], one of the top 10 animal sound collections in the world [26]. Our experiments identified geographic anomalies for 12% of 1037 distinct species in the database, with a total of 371 records out of 7049 records. These anomalies were reviewed by biologists and classified into four categories: A) metadata error; B) outdated metadata; C) errors in the distribution range maps and D) possible new species pattern detected. As will be seen, the latter class of errors can feed all kinds of biodiversity analyses – e.g., detecting animal migration due to change in environmental conditions, such as those caused by climate change.

## II. Background and Related Work

### A. Animal Sound Collections

In biodiversity studies, there is growing interest in sound recordings. Several organizations around the world maintain extensive animal sound collections, providing information not only about species but also about the environment in which they live. These collections have priceless historical information that can be used, for example, to study animal sound communication and behavior – e.g. animals' use of their acoustic and vibrational senses to detect the presence of both predators and prey and to communicate with members of the same species [13].

In addition to the sound recordings, these collections often provide information related to the environment where the sound was recorded, e.g. weather conditions. Such information is widely used in animal habitat prediction, detection of spatial patterns, dynamics of populations, animal conservation, and so on. This helps scientists derive correlations about species, simulate habitat conditions, and conduct countless other studies that help elucidate the past, describe the present and study the future of eco diversity.

Our case study, FNJV, has recordings of all vertebrate groups (fishes, amphibians, reptiles, birds and mammals) and some groups of invertebrates (as insects and arachnids). Other sound recording collections exist as well. The Cornell Lab of Ornithology [10] is an international center for the study, appreciation, and conservation of birds [6]. Fonozoo [27], Fonoteca Zoológica, is yet another example of sound collections, being the animal sound library of the Museo Nacional de Ciencias Naturales of Madrid (Spain) [27]. Currently Fonozoo provides about 33,000 metadata records online. The Animal Sound Archive [15] at the Museum fur Naturkunde in Berlin presently provides about 120,000 bioacoustical recordings. The Avian Knowledge Network [1] provides data from bird-monitoring, bird-banding, and broad-scale citizen-based bird-surveillance programs.

Such collections differ primarily in their number of recordings, the kind of species they have recorded and methods used to obtain recordings. Most of those collections have associated metadata. Such metadata may differ, but the most important fields are supported by all (i.e., recording "what (species observed), when, where, who (observer)"). For example, at FNJV most of the sounds are recorded by domain experts, who often annotated associated metadata during recording time. On the other hand, some of the collections cited above have most of the sounds provided by volunteers. Therefore, in the latter case, there is no quality control of the metadata provided, and thus curation requires additional procedures.

Even though there is no consensual standard on defining metadata fields for sound records, most of them have a common subset of fields. Table I shows 22 (out of 51) metadata fields that are present in the FNJV collection. Row 1 gives information to identify the recorded species (what). Row 2 describes when, where and the environment in which the sound was recorded. Row 3 describes the recording features, as well as devices used to record them (how).

TABLE I. Subset of metadata fields of the FNJV collection.

| | METADATA FIELD |
|---|---|
| 1 | Phylum, Class, Order, Family, Genus, Species, Gender, Number of individuals. |
| 2 | Collect time, Collect date, Country, State, City, Location, Habitat, Micro-Habitat, Air temperature (°C), Atmospheric conditions. |
| 3 | Recording device, Microphone model, Microphone model, Sound file format, Frequency (kHz). |

### B. Geographic Distribution Maps

Citizen science is the term often used to describe communities or networks of citizens who act as observers in some domain of science. Analogously, Volunteer Geographic Information (VGI) refers to data provided by citizens, in particular, including geographic information [20] [16].

Several kinds of projects take advantage of information provided by citizen science and VGI. One example is [24], where citizens measure their personal exposure to noise in their everyday environment by using GPS-equipped mobile phones as noise sensors. The information is used to provide geographic distribution maps about noise-pollution. Such maps can be used to support insight into the problem of urban noise pollution and its social implications.

Another example is the Christmas Bird Count [23]. It is related to animal preservation and environmental studies (our case study). This project is an effort to perform a mid-winter census of bird populations. This kind of project considers, among others, information provided by citizen science and

VGI to create geographic distribution range maps for several species.

Geographic distribution maps are used to show spatial distribution in several domains, e.g., occurrence of diseases, crimes, accidents, species habitat, and so on. Species distribution maps – nowadays more often in digital formats – are commonly used by biologists in their studies. Some maps provide geographic distribution for both current and extinct species, such as the BirdLife International Digital Distribution [5] and the International Union for Conservation of Nature (IUCN) [21].

Distribution maps are usually computed from the combination of a variety of sources, including: a) museum data; b) distribution atlases derived from systematic surveys; c) expert opinions and research expeditions and d) observation records provided by volunteers (citizen science and VGI). The accuracy of these maps can be affected by the quality of the data (especially when provided by non-expert volunteers). As a result, the maps may underestimate/overestimate geographic distribution ranges. Nevertheless, they remain an excellent source of information for several kinds of research.

### C. Incomplete Metadata and Uncertain/Imprecise Location

We find it useful to classify methods for cleaning and curating of observation data as either non-geographic or geographic-based. This classification is focused on domains in which location metadata plays an important role (e.g., environmental studies, epidemiology or biodiversity). We call these domains "location-sensitive," in the sense that geographic information is key for a wide range of scientific analyses.

In a non-geographic-based approach, metadata quality improvement does not consider geographic information present in the metadata as a source of clues for detecting errors. For example, in a manual curation process, biologists may listen to species vocalizations in order to verify if species were correctly identified, but their analysis may not consider the location where the observation was performed. Other examples concern computerized approaches, such as [4], [22], [12] and [2]. In [4], the authors detect duplicated records in metadata using text distance functions. In [22], the authors use clustering methods and association rules in order to perform data cleaning. In [12], authors improve the quality of relational data using conditional functional dependencies. In [2], the authors created a framework that provides metrics to evaluate the expertise of the users and the reliability of data provided by them.

However, some errors can only be detected if the approach considers the location in which the observation occurred. Consider, for instance, metadata that indicate that a polar bear was observed in the Southern hemisphere. A non-geographic approach could not detect that there is an error in the metadata, since polar bears live in the Northern hemisphere.

Geographic-based approaches consider location metadata. The older the observation metadata are, the higher the chance that place information is not georeferenced, and that just location names appear. Even when names are provided, it is not uncommon for the metadata to be incomplete. Uncertain or imprecise descriptions of locations are recurrent problems in observation databases, as are old place names, or references to places that no longer exist. The basic idea, in this case, is to design algorithms that derive coordinate information from place names [7]. In [29], for example, the authors developed filters to improve the quality of data provided by citizen science. Such filters, among other features, take into account the location where species are expected to live, in order to find species that have been misidentified by users. However, this approach does not deal with uncertain and imprecise descriptions of locations, nor can it detect outdated species names in legacy collections.

Indeed, it is not unusual for metadata to be incomplete in biological observation databases, in particular legacy collections. In some cases, missing information, such as air temperature and rainfall indexes, can be derived from external data sources, as we have shown previously [11]. This kind of information can be derived taking into account both the date and location in which an observation was made.

In legacy observation databases, before the GPS era, location information was provided as textual description of the places where recordings were made, e.g. Campo Grande (city), Mato Grosso do Sul (state), Brazil. In this example, deriving the city's centroid coordinates from text does not pose big challenges, since currently there are several techniques to extract this information from gazetteers [17]. Centroid-based approaches, however, may fail to provide precision in the degree needed.

Location information can also be incomplete or imprecise, e.g., some records give only the country names, with no clue about a more specific location. Location metadata may also be recorded as "Brazil, Argentina" because the observation was performed somewhere on the border. Geographic-based cleaning methods must deal with this issue.

## III. Our Approach

The main idea behind our geographical approach is to contrast geographic distribution maps against the places where the observations were made (as per location metadata). When this analysis detects that some of the location observations are not within the expected distribution region, then there is a problem to solve. For example, the metadata are incorrect, or the distribution map presents inconsistencies, etc. The records where problems were identified are then flagged, so scientists can feed the results to subsequent analysis processes.

Our technique can be used with any kind of location-aware observation (e.g. observations about animals, diseases, plants and people), contrasted against the geographic distribution maps of such observations. For example, consider metadata containing locations where people contracted Dengue fever in Brazil (Dengue fever is a disease transmitted mainly through the *Aedes aegypti* mosquito). In this example, the metadata can be contrasted against some authoritative map about this disease, e.g., provided by the Brazilian Ministry of Health, to detect inconsistencies. Without loss of generality, in order to clarify our explanations, this section describes our technique as applied to the domain of animal sound observations (our case study). Figure 2 gives an overview of our approach.

**Step 1 - Preprocessing** The first step of our technique (preprocessing – item C) retrieves from metadata both species

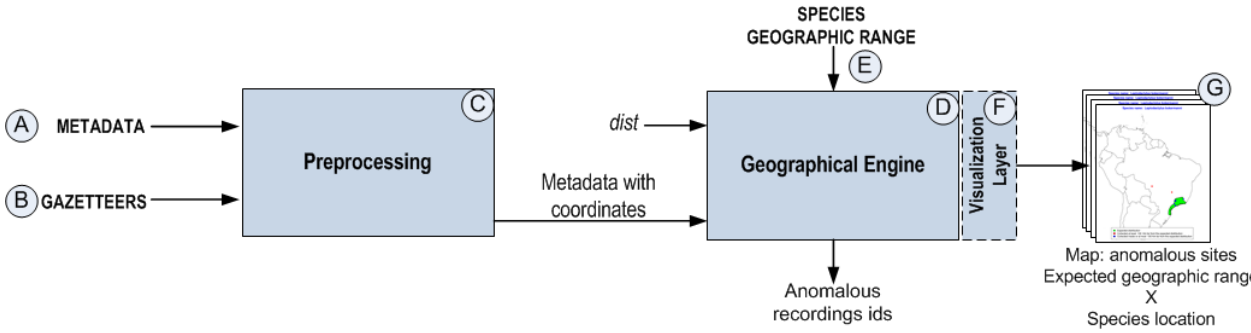


Fig. 2. Geographical technique to support metadata quality improvement in biological observation databases.

name $s$ and the set of places where the species were observed, $P_s$={$p_1$, $p_2$, ..., $p_n$}, where $p_i$ is a point or a polygon that refers to the geographic coordinates of observation record $i$, and $n$ is the number of observations of species $s$ (and thus the number of metadata records for species $s$). The older the metadata information, the higher is the chance that places are not georeferenced, and that just location names appear. Since geographic coordinates are a key aspect in our technique, the preprocessing step also provides a functionality to derive geographic coordinates from gazetteers (item B). The coordinates are obtained in two distinct scenarios: (a) complete location metadata are provided, such as <country, state, city>; (b) incomplete location metadata are provided, such as <country, state>or <country>. In scenario (a), gazetteers provide point coordinates, representing the city's centroid – (City/county names are often used in location metadata, to denote the closest region in which a species was observed) – while scenario (b) provides polygon coordinates of the region indicated in the metadata. Figure 3 shows a map with coordinates for two places, extracted from gazetteers. *Place 1* is a point that represents the city of Campinas, São Paulo state, Brazil (complete metadata location). *Place 2* is a polygon that represents the Brazilian state of Mato Grosso do Sul (incomplete metadata location).

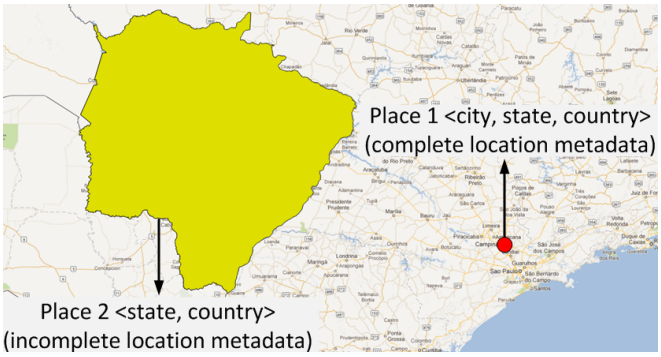


Fig. 3. Coordinates of two places extracted from gazetteers. Place 1 (a point) is derived from a complete metadata location, containing <city, state, country>names. Place 2 (a polygon) is derived from an incomplete metadata location, containing <state, country>names. *(Best in color)*

**Step 2 - Finding anomalous places.** Once the appropriate coordinates are defined, the preprocessing step delivers the *metadata with coordinates* to be processed by the *Geographical Engine* (item D), the core of our approach. It collects data from authoritative geographic distribution maps and processes them against stored metadata, finding anomalous locations as follows.

First, this step retrieves the geographic range (item E) where the species $s$ is expected to live, $GR_s$={$q_1$, $q_2$, ..., $q_m$}, where $GR_s$ is a set of polygons $q$ and $m \geq 1$. $GR_s$ can be retrieved from sources such as the International Union for Conservation of Nature (IUCN) [21] or BirdLife International Digital Distribution Maps of Birds [5]. Although they are authoritative organizations, the regions reported by these kinds of sources are not highly accurate and are known in some cases to be underestimated or overestimated (as explained in section II-B). Furthermore, a domain expert may consider that an observation just a few kilometers beyond $GR_s$ is not an anomaly. In order to overcome this issue, the technique defines a buffered geographic range for $s$, $BGR_s$. It is based on the configuration variable *dist* (one of the inputs of the *geographical engine* step) defined by a domain expert.

$$BGR_s = GR_s + Buffer(dist, GR_s)$$

$BGR_s$ expands the original geographic range $GR_s$ up to its buffer of size *dist*, $Buffer(dist, GR_s)$. Figure 4 shows the original geographic range $GR_s$, the *dist* variable set up by the expert, the buffer and the new buffered region $BGR_s$.

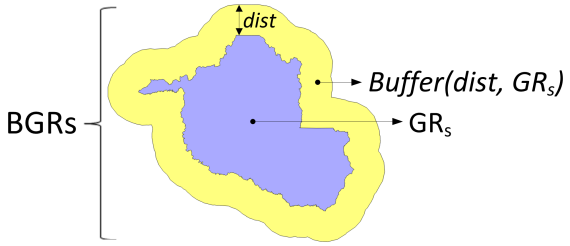


Fig. 4. A buffered geographic range, BGR$_s$, of size *dist*.

Note that if the domain expert considers that $GR_s$ is overestimated, he or she can define a negative value for *dist*, in order to shrink the region provided by the species geographic distribution map. In this case, $BGR_s$ is going to be smaller than $GR_s$. Also note that the variable *dist* may have different values for different kinds of observations. For example, the domain expert may want to set up a higher *dist* value for a specific mammal species than for amphibians because some kinds of mammals can easily move to farther regions.

Given these definitions, anomalous places are defined to

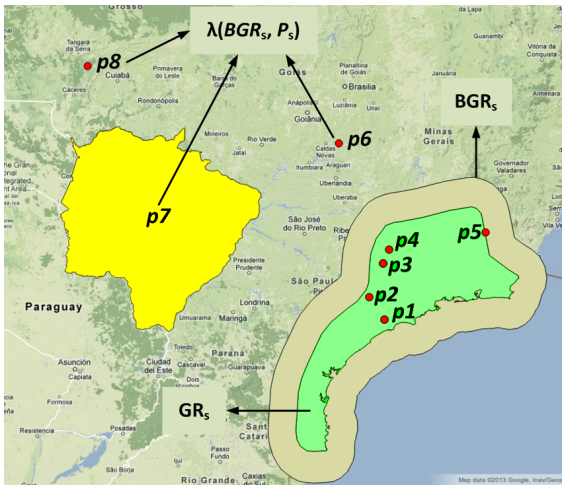


Fig. 5. Example applied in the technique. The green polygon is *GRs* and the gray polygon is *BGRs*. Places from $p_1$ to $p_8$ are the places in which the vocalizations of species *Leptodactylus bokermanni* were recorded, $P_s$. The places $p_6$, $p_7$ (yellow region) and $p_8$ are anomalous, i.e. $\lambda_{(BGRs,Ps)} = \{p_6, p_7, p_8\}$. *(Best in color)*

be elements of $P_s$ that fall outside or do not intersect $BGR_s$. Spatial operations include methods to detect if a set of spatial elements (points and polygons) are *inside* or *intersect* polygons and to calculate the buffer area. The anomalous places are defined as follows:

$$\lambda_{(BGRs,Ps)} = P_s - (P_s \cap BGR_s)$$

The intersection symbol in the definition above retrieves spatial objects as follows. As $P_s$ may contain points and polygons, and $BGR_s$ contains only polygons, then the intersection operation must detect point *in* polygon and polygon *overlapping* polygon. The intersection result is then subtracted from Ps, such that $\lambda_{(BGRs,Ps)} \subseteq P_s$.

**Step 3 - Presenting output to the experts.** The *Geographical Engine* (item D) then delivers information to the visualization layer (item F). This layer creates maps (item G), portraying $P_s$ elements (anomalous and non anomalous) and $BGR_s$ regions. Results provided by the technique comprise such maps and also a list of metadata record ids.

Let us illustrate the process with an example. Consider an animal sound database with 8 recordings of the species *Leptodactylus bokermanni*, a kind of frog. Figure 5 shows the places in which the vocalizations were recorded, $P_s = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8\}$. Note that $p_7$ is a polygon, meaning that the metadata location information for this recording are incomplete (only <state, country>was reported). The green region is the expected geographic range, $GR_s$, for such species. The gray region is calculated based on the variable *dist* set up by the expert. Both gray and green regions comprise $BGR_s$. In this example, the technique singles out vocalizations recorded outside $BGR_s$, i.e., $\lambda_{(BGRs,Ps)} = \{p_6, p_7, p_8\}$.

Given the outputs of Step 3, the scientist can then analyze the results provided. If they show, for example, that a species was observed outside $BGR_s$, the expert can check the data and verify, for instance, if the species was misidentified. If it was, the expert detected an error in the database and can fix it. If it was not, the expert can investigate if it is a new behavior and/or pattern, or even an error in the maps. The classification of the results is performed manually by the domain experts.

Our approach is suitable to any kind of scientific, location-sensitive metadata database, especially large collections. It provides support to tasks that would not be possible to perform manually in an acceptable time frame. It is important to note that the process flow does not define if the data are wrong or if a pattern was detected. The process is semi-automatic, being used to help experts to improve metadata quality.

Our technique proved to be also useful, among other things, to perform detection of outdated records, as described in the case study detailed in the next section.

## IV. Case Study

### A. Data Preparation

Our case study addressed the needs of curators of FNJV. The original collection dates back to the 1960's, and thus most records lack geographic coordinates of where sounds were recorded, $P_s$. Instead, there is an indication of place names. First, we derived missing coordinates from Geonames [28] and the Brazilian Institute of Geography and Statistics (IBGE) [18], using centroids of polygons (cities, states, countries). Note that this methodology may not provide accurate coordinates of the places where the sounds were recorded, However, this approximation was deemed by the experts to be good enough for the purposes of our case study (as confirmed in the subsequent tests).

For the species spatial distribution maps, $GR_s$, we downloaded shapefile files provided by the IUCN Red List [21] and the BirdLife International Digital Distribution Maps of Birds [5]. These files were adjusted to the EPSG 4326 geographic coordinate system and WGS84 world geodetic system. Additional map sources can also be used (e.g., National Atlas Amphibian distribution [3]).

### B. Prototype

Our prototype was created using R [19], a language and environment for statistical computing and graphics. We chose R because it provides a wide variety of statistical and graphical techniques as well as because it is highly extensible. Figure 6 presents the architecture of the prototype. It has four inputs: 1) animal sound collection data (item 1) in which $P_s$ (place coordinates – points and polygons) are provided in SHP format; 2) species geographic range maps (items 4 and 5), $GR_s$ (polygons in SHP format); 3) the *dist* parameter (item 6); and 4) the place coordinates provided by IBGE and Geonames (items 2 and 3). In particular, IBGE data were provided in KML file format, and Geonames data were provided through web services. Coordinates were extracted from IBGE KML files using Java JDOM API. Geonames web service was accessed using a JAVA API provided by Geonames (coordinates were retrieved through the API functions). Coordinates were saved in the FNJV database and then exported to the SHP file format.

Our prototype provides two outputs: 1) Textual description: a list of database records (including the row id) in which species vocalizations were recorded out of the $BGR_s$; 2) Visual description: maps containing the rows id, species name, the

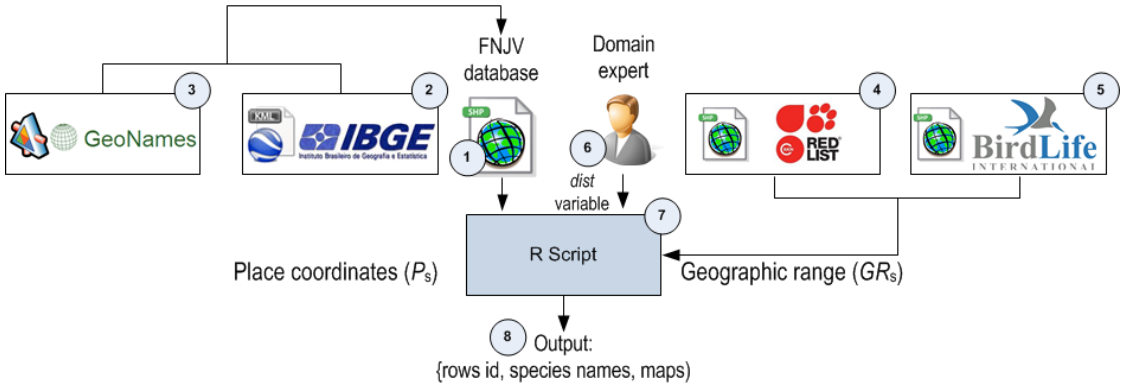


Fig. 6. Prototype of our Geographical Approach for Metadata Quality Improvement.

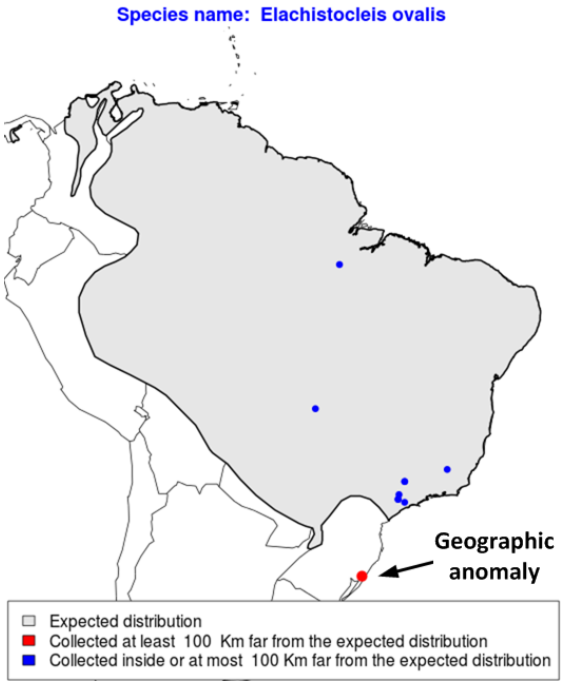


Fig. 7. Output map generated by the prototype for the *Elachistocleis ovalis* species (Amphibian)– anomaly classified as *outdated metadata*. *(Best in color)*

regions $GR_s$ where they are expected to live and the places $P_s$ where the vocalizations were recorded.

Figures 7 and 8 show two output maps generated by our prototype (the maps exhibit part of South America). These maps refer to the *Elachistocleis ovalis* and *Allobates marchesianus* species. The gray polygons represent the regions in which these species are expected to live, $GR_s$ (according to IUCN). Points represent the places where the species sounds were recorded, $P_s$. Points are colored blue when *inside* $BGR_s$. They are colored red beyond the *dist* tolerance, i.e., red points $\in \lambda_{(BGRs,Ps)}$.

In contrast, Figure 9 shows a map for the *Aplastodiscus perviridis* species. This map shows that all observations of such species were made *inside* $BGR_s$, i.e., all observations were non-anomalous.

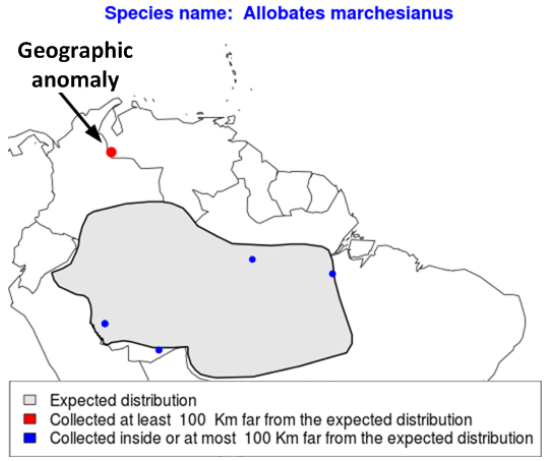


Fig. 8. Output map generated by the prototype for the *Allobates marchesianus* (Amphibian) – anomaly classified as *error in the distribution range map*. *(Best in color)*

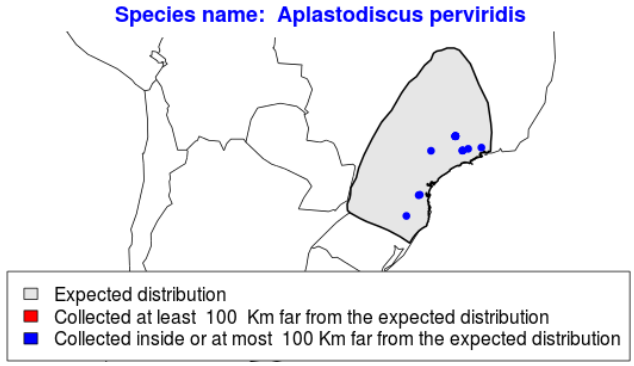


Fig. 9. Output map generated by the prototype for the *Aplastodiscus perviridis* species, an amphibian species. In this case all observations are non-anomalous. *(Best in color)*

### C. Results

The prototype was set up with *dist* = 100 kilometers. Table II summarizes some of the input and output numbers of our experiment. The first column shows the four distinct classes of animals we used as input: Bird, Mammal, Reptile and Amphibian. The second column shows the number of observations for each taxonomic class. The third column shows

the number of observations which were detected in anomalous places. The fourth column describes the number of distinct species analyzed. The last column shows the number of distinct species which were detected in anomalous places. Among 1037 distinct species in our case study (119 Amphibians, 877 Birds, 38 Mammals and 3 Reptiles), 13 Amphibian, 105 Birds, 11 Mammals and 0 Reptiles species were detected in anomalous sites, i.e., about 12%.

TABLE II. DETAILS FOR EACH SPECIES CLASS USED IN OUR EXPERIMENT. COMPARISON OF THE NUMBER OF RECORDS/SPECIES ANALYZED AND THE NUMBER OF ANOMALIES DETECTED.

| Species Class | Observations in the sound database | Anomalous observations | Species analyzed | Species detected in anomalous places |
|---|---|---|---|---|
| Amphibian | 419 | 21 | 119 | 13 |
| Bird | 6414 | 303 | 877 | 105 |
| Mammal | 212 | 47 | 38 | 11 |
| Reptile | 4 | 0 | 3 | 0 |

The maps with anomalous places generated by our prototype were presented to biologists, who manually classified the anomalies into 4 categories: A) metadata errors; B) outdated metadata; C) errors in distribution range maps and D) anomalous pattern. Table III details each category.

TABLE III. CLASSIFICATION OF THE EXPERIMENT RESULTS INTO FOUR CATEGORIES.

| Class | Classification | Description |
|---|---|---|
| A | Metadata error | species were wrongly classified (biologists must listen to the recordings in order to correctly reclassify the species) |
| B | Outdated metadata | the scientific name changed (biologists must verify current taxonomic information and update the metadata) |
| C | Errors in the distribution range maps | species geographic range maps may be overestimated or underestimated |
| D | Anomalous pattern | new distributional record for the species, this may promote advances in our understanding of animal distribution. Scientists must use data mining methods to detect the cause of the anomalous pattern. |

Figure 7 shows an output map generated by our prototype for the amphibian species *Elachistocleis ovalis*. Scientists informed that the species taxonomic name was divided into several other species. This anomaly corresponds to outdated metadata (Table III, Class B). Figure 8 shows a map for the species *Allobates marchesianus*. According to the domain expert, the species distribution range map is underestimated. This anomaly corresponds to an error in distribution range map (Table III, Class C).

Table IV shows four kinds of feedback from scientists about the amphibian species detected in anomalous sites. The first column gives the names of the species in the metadata. The second column shows the corresponding number of anomalous database records. The third column describes the category in which the scientist classified the anomaly (according to Table III). The fourth column summarizes the corresponding feedback.

Let us clarify the content of Table IV by detailing the fourth row (*Pseudis limellum* species). For such species, six database records had vocalizations recorded in anomalous places. Scientists analyzed such records, concluding that the sounds recorded in the Amazon forest region probably correspond to other species (probably *Lysapsus limellum*). The

TABLE IV. BIOLOGISTS FEEDBACK FOR AMPHIBIAN SPECIES OBSERVED IN ANOMALOUS SITES.

| Species name | Anomalous observations in the database | Class | Comments by Scientists |
|---|---|---|---|
| *Allobates marchesianus* | 1 | C | Probably the distribution map is underestimated. |
| *Elachistocleis ovalis* | 1 | B | This name is not valid any longer. Species subdivided into others (all of which with smaller spatial distribution than their predecessor) |
| *Leptodactylus bokermanni* | 2 | A or D | The points in the middle of Brazil probably are other species. They might be the *Adenomera bokermanni*. |
| *Pseudis limellum* | 6 | A or D | The point in the Amazon forest region probably corresponds to other species, perhaps *Lysapsus limellum*. |

anomalies were classified as categories A or D (metadata error or new pattern – according to Table III). It means that at first glance the animal sounds were misidentified (metadata error). However, if the domain expert double checks the recording of such records and verifies that the species was correctly identified, the anomaly is actually a new pattern of species distribution, with important implications in biodiversity studies. For instance, since the study involves legacy data, this may indicate that species migrated from that region (and thus it is up to the experts to analyze historical records on that same region to see what changed to cause such migration). In some cases, such records may be the only witness to the fact that the species actually lived in that area.

Another interesting fact from Table IV comes from its first row. This example indicates that the range maps provided by authoritative sources may be wrong. Thus, not only can we detect errors in metadata, but indicate problems with consensual external sources.

## V. CONCLUSIONS AND FUTURE WORK

We presented a geographical technique to improve metadata quality in biological observation databases for domains in which location plays an important role. Our experiment results were manually analyzed by domain experts, who classified the results into four categories: metadata errors, outdated metadata, errors in species distribution range maps and possible new species pattern. Our work has been motivated by challenges faced by biologists on managing large amounts of animal sound recordings, using FNJV as a real world case study.

Ongoing and future work might focus on employing supervised learning algorithms to recommend classes to be reviewed by scientists, to reduce their burden. We also intend to consider environmental variables in our approach, by using enriched information provided by our previous work [11].

## ACKNOWLEDGMENT

We thank Prof. Scott Lanyon and Milena Corbo for their insightful comments. We also thank the University of Minnesota Spatial Databases and Spatial Data Mining Research Group for their comments, and Kim Koffolt, who improved the paper's readability. Work partially financed by FAPESP (grants 2011/19284-3, 2012/11395-3, 2008/50325-5 and 2011/51694-7), the Microsoft Research FAPESP Virtual Institute (NavSCales project), CNPq (MuZOO Project), INCT in Web Science,

CAPES, CNPq, NSF (grant 1029711) and USDOD (grants HM1582-08-1-0017 and HM1582-07-1-2035).

## REFERENCES


[1] AKN. Avian knowledge network. http://www.avianknowledge.net (Accessed on 12/2012).

[2] A. Alabri and J. Hunter. Enhancing the quality and trust of citizen science data. In *IEEE VI International Conference one-Science*, pages 81–88. IEEE, 2010.

[3] N. Atlas. Amphibians distribution. http://www.nationalatlas.gov/mld/amphibt.html (Accessed on: January, 2013).

[4] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, New York, NY, USA, 2003. ACM.

[5] BirdLife International. Digital Distribution Maps of the Birds of the Western Hemisphere, version 5.0. BirdLife International and NatureServe, 2012.

[6] R. Bonney. Citizen science at the cornell lab of ornithology. *Exemplary Science in Informal Education Settings: Standards-based Success Stories*, pages 213–229, 2007.

[7] K. A. Borges, A. H. Laender, C. B. Medeiros, and C. A. Davis Jr. Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 31–36. ACM, 2007.

[8] S. Bowers, J. Kudo, H. Cao, and M. P. Schildhauer. Obsdb: A system for uniformly storing and querying heterogeneous observational data. In *IEEE Sixth International Conference on e-Science*, pages 261–268. IEEE, 2010.

[9] A. Chapman. Principles of data quality. *Report for the Global Biodiversity Information Facility. Copenhagen, Denmark.*, pages 1–58, 2005.

[10] Cornell. The cornell lab of ornithology. http://www.allaboutbirds.org (Accessed on 12/2012).

[11] D. C. Cugler, C. B. Medeiros, and L. F. Toledo. An architecture for retrieval of animal sound recordings based on context variables. *Concurrency and Computation: Practice and Experience*, June 2012.

[12] W. Fan, F. Geerts, and X. Jia. Semandaq: a data quality system based on conditional functional dependencies. *Proceedings of the VLDB Endowment*, 1(2):1460–1463, 2008.

[13] N. Fletcher. Animal bioacoustics. *Springer Handbook of Acoustics, ISBN 978-0-387-30446-5. Springer-Verlag New York, 2007, p. 785*, 1:785, 2007.

[14] FNJV. Online animal sound collection - Fonoteca Neotropical Jacques Vielliard. http://proj.lis.ic.unicamp.br/fnjv (Accessed on 06/2013).

[15] K.-H. Frommolt, R. Bardeli, F. Kurth, and M. Clausen. The animal sound archive at the humboldt-university of berlin: Current activities in conservation and improving access for bioacoustic research. In *Advances in Bioacoustics II*, pages 139–144, 2006.

[16] M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007.

[17] L. L. Hill. *Georeferencing: The geographic associations of information*. MIT Press, 2009.

[18] IBGE. Brazilian institute of geography and statistics. http://www.ibge.gov.br (Accessed on 05/2013).

[19] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[20] A. Irwin. *Citizen science: A Study of People, Expertise, and Sustainable Development*, volume 136. Routledge Londres, 1995.

[21] IUCN International Union for Conservation of Nature. IUCN Red List of Threatened Species, 2012.

[22] R. Kumar and R. Chadrasekaran. Attribute correction–data cleaning using association rule and clustering methods. *Intl. Jrnl. of Data Mining & Knowledge Management Process*, 1(2):22–32, 2011.

[23] G. S. LeBaron, R. J. Cannings, D. K. Niven, G. S. Butcher, G. T. Bancroft, P. W. Sykes Jr, S. M. Elliott, N. Strycker, and P. Read. The 109th christmas bird count. *American Birds*, 63:2–7, 2009.

[24] N. Maisonneuve, M. Stevens, M. E. Niessen, P. Hanappe, and L. Steels. Citizen noise pollution monitoring. In *10th Int. Conference on Digital Government Research*, pages 96–103. Digital Government Society of North America, 2009.

[25] E. Rahm and H. Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000.

[26] R. Ranft. Natural sound archives: past, present and future. *Anais da Academia Brasileira de Ciências*, 76(2):455–465, 2004.

[27] G. Solís, X. Eekhout, and R. Márquez. Fonoteca zoologica (www. fonozoo. com): the web-based animal sound library of the museo nacional de ciencias naturales (madrid), a resource for the study of anuran sounds. In *Proceedings of the 13th Congress of the Societas Europaea Herpetologica. pp*, volume 171, page 174, 2006.

[28] M. Wick. Geonames. http://www.geonames.org/ (Accessed on 12/2012).

[29] J. Yu, S. Kelling, J. Gerbracht, and W.-K. Wong. Automated data verification in a large-scale citizen science project: A case study. In *IEEE VIII International Conference on E-Science*, pages 1–8. IEEE, 2012.