

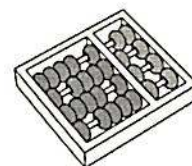
Daniel Cintra Cugler

**“Supporting the Collection and Curation of  
Biological Observation Metadata”**

***“Apoio à Coleta e Curadoria de Metadados de  
Observações Biológicas”***

**CAMPINAS  
2014**





University of Campinas  
Institute of Computing

*Universidade Estadual de Campinas  
Instituto de Computação*

Daniel Cintra Cugler

## “Supporting the Collection and Curation of Biological Observation Metadata”

Supervisor: Prof<sup>a</sup>. Dr<sup>a</sup>. Claudia Bauzer Medeiros  
*Orientador(a):*

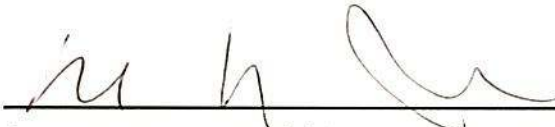
### *“Apoio à Coleta e Curadoria de Metadados de Observações Biológicas”*

PhD Thesis presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a PhD degree in Computer Science.

*Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Doutor em Ciência da Computação.*

THIS VOLUME CORRESPONDS TO THE VERSION OF THE THESIS SUBMITTED TO EXAMINING BOARD BY DANIEL CINTRA CUGLER, UNDER THE SUPERVISION OF PROF<sup>A</sup>. DR<sup>A</sup>. CLAUDIA BAUZER MEDEIROS.

*ESTE EXEMPLAR CORRESPONDE À VERSÃO DA TESE APRESENTADA À BANCA EXAMINADORA POR DANIEL CINTRA CUGLER, SOB ORIENTAÇÃO DE PROF<sup>A</sup>. DR<sup>A</sup>. CLAUDIA BAUZER MEDEIROS.*

  
Supervisor's signature / *Assinatura do Orientador(a)*

CAMPINAS  
2014

Ficha catalográfica  
Universidade Estadual de Campinas  
Biblioteca do Instituto de Matemática, Estatística e Computação Científica  
Maria Fabiana Bezerra Muller - CRB 8/6162

C895s Cugler, Daniel Cintra, 1982-  
Supporting the collection and curation of biological observation metadata /  
Daniel Cintra Cugler. – Campinas, SP : [s.n.], 2014.

Orientador: Claudia Maria Bauzer Medeiros.  
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de  
Computação.

1. Ciência da computação. 2. Banco de dados. 3. Limpeza de dados. 4.  
Biologia - Banco de dados. I. Medeiros, Claudia Maria Bauzer, 1954-. II.  
Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

**Título em outro idioma:** Apoio à coleta e curadoria de metadados de observações biológicas

**Palavras-chave em inglês:**

Computer science

Databases

Data cleaning

Biology - Databases

**Área de concentração:** Ciência da Computação

**Titulação:** Doutor em Ciência da Computação

**Banca examinadora:**

Claudia Maria Bauzer Medeiros [Orientador]

Clodoveu Augusto Davis Junior

Luciano Antonio Digiampietri

André Santanchè

Fabio Luiz Usberti

**Data de defesa:** 01-09-2014

**Programa de Pós-Graduação:** Ciência da Computação

## TERMO DE APROVAÇÃO

Defesa de Tese de Doutorado em Ciência da Computação, apresentada pelo(a)  
Doutorando(a) **Daniel Cintra Cugler**, aprovado(a) em **1º de setembro de 2014**, pela  
Banca examinadora composta pelos Professores Doutores:

  
**Prof<sup>(a)</sup>. Dr<sup>(a)</sup>. Clodoveu Augusto Davis Junior**  
**Titular**

  
**Prof<sup>(a)</sup>. Dr<sup>(a)</sup>. Luciano Antonio Digiampietri**  
**Titular**

  
**Prof<sup>(a)</sup>. Dr<sup>(a)</sup>. André Santanchè**  
**Titular**

  
**Prof<sup>(a)</sup>. Dr<sup>(a)</sup>. Fabio Luiz Usberti**  
**Titular**

  
**Prof<sup>(a)</sup>. Dr<sup>(a)</sup>. Claudia Maria Bauzer Medeiros**  
**Presidente(a)**



# Supporting the Collection and Curation of Biological Observation Metadata

Daniel Cintra Cugler<sup>1</sup>

September 01, 2014

## Examiner Board / *Banca Examinadora*:

- Prof<sup>a</sup>. Dr<sup>a</sup>. Claudia Maria Bauzer Medeiros (Supervisor / *Orientadora*)
- Prof. Dr. André Santanchè  
Institute of Computing - UNICAMP
- Prof. Dr. Fábio Luiz Usberti  
Institute of Computing - UNICAMP
- Prof. Dr. Clodoveu Augusto Davis Junior  
DCC - UFMG
- Prof. Dr. Luciano Antonio Digiampietri  
EACH - USP

---

<sup>1</sup>Financial support: Work partially financed by FAPESP (grants 2011/19284-3 and 2012/11395-3), CAPES, FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and individual grants from CNPq.





# Abstract

Biological observation databases contain information about the occurrence of an organism or set of organisms detected at a given place and time according to some methodology. Such databases store a variety of data, at multiple spatial and temporal scales, including images, maps, sounds, texts and so on. This priceless information can be used in a wide range of research initiatives, e.g., global warming, species behavior or food production. All such studies are based on analyzing the records themselves, and their metadata. Most times, analyses start from metadata, often used to index the observation records. However, given the nature of observation activities, metadata may suffer from quality problems, hampering such analyses. For example, there may be metadata gaps (e.g., missing attributes, or insufficient records). This can have serious effects: in biodiversity studies, for instance, metadata problems regarding a single species can affect the understanding not just of the species, but of wider ecological interactions.

This thesis proposes a set of processes to help solve problems in metadata quality. While previous approaches concern one given aspect of the problem, the thesis provides an architecture and algorithms that encompass the whole cycle of managing biological observation metadata, which goes from acquiring data to retrieving database records. Our contributions are divided into two categories: (a) data enrichment and (b) data cleaning. Contributions in category (a) provide additional information for both missing attributes in existent records, and missing records for specific requirements. Our strategies use authoritative remote data sources and VGI (Volunteered Geographic Information) to enrich such metadata, providing missing information. Contributions in category (b) detect anomalies in biological observation metadata by performing spatial analyses that contrast location of the observations with authoritative geographic distribution maps. Thus, the main contributions are: (i) an architecture to retrieve biological observation records, which derives missing attributes by using external data sources; (ii) a geographical approach for anomaly detection and (iii) an approach for adaptive acquisition of VGI to fill out metadata gaps, using mobile devices and sensors. These contributions were validated by actual implementations, using as case study the challenges presented by the management of biological observation metadata of the Fonoteca Neotropical Jacques Viellard (FNJV), one of the top 10 animal sound collections in the world.



# Resumo

Bancos de dados de observações biológicas contêm informações sobre ocorrências de um organismo ou um conjunto de organismos detectados em um determinado local e data, de acordo com alguma metodologia. Tais bancos de dados armazenam uma variedade de dados, em múltiplas escalas espaciais e temporais, incluindo imagens, mapas, sons, textos, etc. Estas inestimáveis informações podem ser utilizadas em uma ampla gama de pesquisas, por exemplo, aquecimento global, comportamento de espécies ou produção de alimentos. Todos estes estudos são baseados na análise dos registros e seus respectivos metadados. Na maioria das vezes, análises são iniciadas nos metadados, estes frequentemente utilizados para indexar os registros de observações. No entanto, dada a natureza das atividades de observação, metadados podem possuir problemas de qualidade, dificultando tais análises. Por exemplo, podem haver lacunas nos metadados (por exemplo, atributos faltantes ou registros insuficientes). Isto pode causar sérios problemas: em estudos em biodiversidade, por exemplo, problemas nos metadados relacionados a uma única espécie podem afetar o entendimento não apenas da espécie, mas de amplas interações ecológicas.

Esta tese propõe um conjunto de processos para auxiliar na solução de problemas de qualidade em metadados. Enquanto abordagens anteriores enfocam em um dado aspecto do problema, esta tese provê uma arquitetura e algoritmos que englobam o ciclo completo da gerência de metadados de observações biológicas, que vai desde adquirir dados até recuperar registros na base de dados. Nossas contribuições estão divididas em duas categorias: (a) enriquecimento de dados e (b) limpeza de dados. Contribuições na categoria (a) proveem informação adicional para ambos atributos faltantes em registros existentes e registros faltantes para requisitos específicos. Nossas estratégias usam fontes de dados remotas oficiais e VGI (Volunteered Geographic Information) para enriquecer tais metadados, provendo as informações faltantes. Contribuições na categoria (b) detectam anomalias em metadados de observações biológicas através da execução de análises espaciais que contrastam a localização das observações com mapas oficiais de distribuição geográfica de espécies. Deste modo, as principais contribuições são: (i) uma arquitetura para recuperação de registros de observações biológicas, que deriva atributos faltantes

através do uso de fontes de dados externas; (ii) uma abordagem espacial para detecção de anomalias e (iii) uma abordagem para aquisição adaptativa de VGI para preencher lacunas em metadados, utilizando dispositivos móveis e sensores. Estas contribuições foram validadas através da implementação de protótipos, utilizando como estudo de caso os desafios oriundos do gerenciamento de metadados de observações biológicas da Fonoteca Neotropical Jacques Vielliard (FNJV), uma das 10 maiores coleções de sons de animais do mundo.

# Acknowledgements

I would like to thank several people and organizations, that directly and indirectly supported this research. Thanks to my advisor, professor Claudia, for sharing her precious knowledge, and for opening several doors in my life. Professor André Santanchè, for all hints and for sharing his good mood 365 days a year. Professor Shashi Shekhar, for sharing his knowledge and for opening the doors of his research group to me. Professor Felipe Toledo, for his cooperative work. Professor Mauro Biajiz (*in memoriam*), for giving me the opportunity to enter in the academic world, what changed my life forever.

My parents, Heloisa and Gilberto, for the magnificent education they gave me and for showing me the most important values of life. My sisters, Vanessa and Priscilla, for being the best sisters someone ever had. My wife, Juliana, for supporting me in every single dream I dream; and my daughter, Helena, for being fuel for my life.

Friends from the Laboratory of Information Systems, University of Campinas: Alan, Alessandra, Andréia, Bruno, Luis Theodoro, Luiz Celso, Eduardo, Ivelize, Ivo, Jaudete, Jacqueline, Joana, Jordi, Leandro, Lucas, Matheus, Nielsen, Renato, Sávio and Senra.

Friends from the Spatial Computing Research Group, University of Minnesota: Dev, Emre, KwangSoo, Mike, Reem, Viswanath, Xun, and Zhe.

Funding agencies; FAPESP (grants 2011/19284-3 and 2012/11395-3), CAPES, FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and individual grants from CNPq.

The Federal Institute of Triângulo Mineiro (IFTM), for providing me special student time (according to ordinance number 642 of May 08, 2014).

Finally, I would like to thank God, for showing me the right paths to follow and for giving me more opportunities than I definitely deserve.



# Contents

<b>Abstract</b>	<b>ix</b>
<b>Resumo</b>	<b>xi</b>
<b>Acknowledgements</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview of the problem . . . . .	1
1.2 Organization and Contributions . . . . .	4
<b>2 An Architecture for Retrieval of Animal Sound Recordings Based on Context Variables</b>	<b>9</b>
2.1 Introduction . . . . .	9
2.2 Related Work and Some Challenges . . . . .	11
2.2.1 Sound Retrieval Based on Recording Metadata . . . . .	12
2.2.2 Sound Retrieval Based on Acoustic Similarity . . . . .	13
2.2.3 Sound Retrieval Based on Semantic Similarity . . . . .	14
2.3 Defining Context Variables to Support Sound Retrieval . . . . .	15
2.4 Publishing Animal Sounds on the Web - Case study and prototype . . . . .	17
2.4.1 Public functionalities . . . . .	18
2.4.2 Private functionalities . . . . .	20
2.4.3 Rationale behind the prototype . . . . .	21
2.5 Semantic retrieval of animal sounds based on context analysis . . . . .	22
2.5.1 An architecture for retrieval of animal sounds based on context analysis . . . . .	22
2.5.2 Extracting information from remote sources to derive context variables . . . . .	25
2.5.3 Extracting information from semantic bases in order to improve queries . . . . .	28
2.6 Final Considerations and Ongoing Work . . . . .	29





<b>3</b>	<b>A Geographical Approach for Metadata Quality Improvement in Biological Observation Databases</b>	<b>31</b>
3.1	Introduction . . . . .	31
3.2	Background and Related Work . . . . .	33
3.2.1	Animal Sound Collections . . . . .	33
3.2.2	Geographic Distribution Maps . . . . .	34
3.2.3	Incomplete Metadata and Uncertain/Imprecise Location . . . . .	35
3.3	Our Approach . . . . .	36
3.4	Case Study . . . . .	41
3.4.1	Data Preparation . . . . .	41
3.4.2	Prototype . . . . .	42
3.4.3	Results . . . . .	43
3.5	Discussion . . . . .	46
3.6	Conclusions and Future Work . . . . .	47
<b>4</b>	<b>Adaptive Acquisition of VGI to Fill out Gaps in Biological Observation Metadata</b>	<b>49</b>
4.1	Introduction and Motivation . . . . .	49
4.2	Background and Related Work . . . . .	51
4.2.1	Biological Observation Metadata . . . . .	51
4.2.2	Volunteered Geographic Information (VGI) . . . . .	52
4.2.3	Approaches to collect VGI . . . . .	53
4.3	Our Approach . . . . .	54
4.3.1	Overview . . . . .	54
4.3.2	Detection of Gaps . . . . .	55
4.4	Prototype . . . . .	58
4.4.1	The Arduino Platform . . . . .	58
4.4.2	Defining on the Fly Which Data Fields Must be Acquired . . . . .	59
4.4.3	Example . . . . .	61
4.5	Conclusions and Future Work . . . . .	65
<b>5</b>	<b>Conclusions and Extensions</b>	<b>67</b>
5.1	Main Contributions . . . . .	67
5.2	Extensions . . . . .	68
	<b>Bibliography</b>	<b>71</b>



# List of Tables

2.1	Subset of metadata fields that are present in the FNJV collection and an example of a metadata record. . . . .	30
3.1	Subset of metadata fields of the FNJV collection. . . . .	34
3.2	Details for each species Class used in our experiment. Comparison of the number of records/species analyzed and the number of anomalies detected. . . . .	44
3.3	Classification of the experiment results into four categories. . . . .	45
3.4	Biologists feedback for amphibian species observed in anomalous sites. . . . .	46



# List of Figures

1.1	Overview of how chapters fit their roles in the Acquisition-Curation-Processing cycle. . . . .	3
2.1	Steps involved between collect and analysis processes. This research concerns step (C). . . . .	11
2.2	Four different call types of the same treefrog species ( <i>Scinax fuscomarginatus</i> ), emitted in different social contexts (modified from [90]). . . . .	13
2.3	Partial screenshot of the page that allows any user to navigate through the whole collection of animal sounds. . . . .	19
2.4	Details about metadata of one recording - geographic metadata on recording site. . . . .	20
2.5	Form to perform metadata-based queries . . . . .	20
2.6	Page that shows the records selected by the user and the link to export them into spreadsheets . . . . .	21
2.7	Semantic retrieval of animal sounds based on context analysis. . . . .	23
2.8	Architecture of prototype to derive contextual information - use of special purpose algorithms, external knowledge bases and data sources. . . . .	26
2.9	Prototype to query Freebase, NASA and IRIS services. . . . .	28
3.1	Overview of our geographical approach. . . . .	33
3.2	Geographical technique to support metadata quality improvement in biological observation databases. . . . .	37
3.3	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. . . . .	38
3.4	A buffered geographic range, BGR <sub>s</sub> , of size <i>dist</i> . . . . .	39
3.5	. . . . .	40
3.6	Prototype of our Geographical Approach for Metadata Quality Improvement. . . . .	41



3.7	Output map generated by the prototype for the <i>Elachistocleis ovalis</i> species (Amphibian)– anomaly classified as <i>outdated metadata</i> . . . . .	42
3.8	Output map generated by the prototype for the <i>Allobates marchesianus</i> (Amphibian) – anomaly classified as <i>error in the distribution range map</i> . .	43
3.9	Output map generated by the prototype for the <i>Aplastodiscus perviridis</i> species, an amphibian species. In this case all observations are non-anomalous.	44
4.1	Our approach for adaptive acquisition of VGI to fill out gaps in biological observation metadata. . . . .	55
4.2	Observations $O_n$ provided by VGI. Blue points are observations provided within the expected spatial coverage $S$ . Red points are out of the spatial coverage. . . . .	57
4.3	Scientists remotely configure forms and sensors, defining which attributes to collect. . . . .	60
4.4	Output map generated by the gap detection algorithm for the <i>Elachistocleis ovalis</i> species (Amphibian). The anomaly (red point) led to the spatial gap detection. . . . .	62
4.5	Hardware used to collect <i>Sensor-based VGI</i> . . . . .	63
4.6	Screenshot of our Android application prototype. The form is created on the fly, based on the XML configuration file. . . . .	64





# Chapter 1

## Introduction

### 1.1 Overview of the problem

Biological observation databases provide several kinds of data, at multiple spatial and temporal scales, such as images, maps, sounds, texts and so on. This data contains information about occurrences of an organism or set of organisms detected at a given place and time according to some methodology. Such databases contain priceless information that can be used to provide knowledge for broad kinds of research, e.g., global warming, species behavior or food production. As the number of records grows, so does the difficulty to manage them, presenting challenges to save, retrieve, share and manage the records. Evidently, the quality of research results is directly associated with the quality of the data used. The retrieval of observation records (e.g., from a repository or database) is usually performed using a combination of two kinds of approaches:

- accessing the actual observation *data* (e.g., retrieving animal sound recordings or species photos, by using pattern matching, or mining techniques);
- accessing the observation *metadata*, as means of finding out more about the observations. Such metadata follows a variety of standards, containing at least information on *WHO* (observer), *WHAT* (species observed), *WHERE* (observation location), *WHEN* (observation period) and *HOW* (observation methodology).

This thesis is concerned with challenges posed by the second approach - i.e., handling metadata records. Related work mostly concentrates on parts of this cycle - e.g., [75, 81, 63, 51, 86, 62] are geared towards collecting metadata; [44, 35] on filling blanks and [5, 99, 26, 82] on detecting anomalies. This thesis, instead, covers the full management cycle from metadata acquisition to gap detection, and improvement of such records.

In particular, the thesis proposes solutions to the following problems:

- **Checking anomalies in metadata:** Species names may change as time goes by, invalidating curated records. Records frequently have wrong information, such as misidentified species. Locations where observations are made, sometimes, are out of the expected species geographical distribution (outliers).
- **Filling gaps in metadata:** Records may contain blank attributes due to, for example, lack of equipments to measure environmental variables, such as air temperature and rainfall indexes. Databases may also have an insufficient number of observation records for some specific scenario, e.g., concerning a time period, a region, or even a research methodology.

Detecting *anomalies in metadata* is challenging because of the variety of issues concerning metadata, e.g., misnaming of species, location uncertainty and imprecision concerning where observations were recorded. Related work on these problems (e.g., [15, 65, 36, 5]) is limited because it usually considers a specific aspect. For example, in [15], the authors detect duplicate records in metadata using text distance functions. In [65] the authors use clustering methods and association rules in order to perform data cleaning. In [36], authors improve the quality of relational data using conditional functional dependencies. In [5], the authors created a framework that provides metrics to evaluate the expertise of the users and the reliability of data provided by them. Though all are interesting contributions, some anomalies can only be detected if considering the location in which an observation occurred, and/or expert knowledge. There is no related work that deals with such geographical aspects to consider the anomalies mentioned (e.g., misnaming of species).

Detecting and filling *gaps in metadata* is challenging due to metadata requirements that change as researchers acquire new knowledge about their problems. While several studies are dedicated to filling blank attributes, to the best of our knowledge, there is no automated proposal to dynamically (re)direct observations to provide missing records, even when using VGI<sup>1</sup> [59, 45]. Related work (e.g., [51, 86, 62]) is limited because it does not take this evolution into consideration. For example, in [51], scientists create electronic forms to collect metadata from VGI through web-based and mobile applications. However, such electronic forms are static and the data acquisition strategy does not consider that metadata requirements may change, requiring different information in distinct time frames. In [86] and [62] authors provide approaches that support the creation of personalized questionnaires to collect VGI. However, they do not enable performing dynamic changes in the questionnaires by considering the gaps that may occur in the metadata. Moreover, there are no dynamic changes in the strategy of where to send volunteers.

---

<sup>1</sup>VGI refers to data provided by citizens, in particular, including geographic information.

Given the limitations of related work, the goal of this thesis is to provide a suite of algorithms to solve some of the problems related to the management of biological observation databases. This work was developed within cooperative research between the Laboratory of Information Systems (LIS) and the Fonoteca Neotropical Jacques Vieliard (FNJV), one of the 10 largest animal sound collections in the world. This research falls within eScience [53], defined as joint research in computer science with scientists of other domains, envisaging the development of solutions that support these scientists on performing their research faster, better or in a different way.

The thesis thus provides algorithms that comprise the whole cycle Acquisition-Curation-Processing in the management of biological observation metadata. Figure 1.1 briefly describes the role of each chapter in this cycle, as well as gives an overview of how chapters are linked. Chapter 4 is responsible for *acquisition* of observations records. Both Chapters 2 and 3 are responsible for *curating* the biological observation database by filling out missing attributes and detecting anomalous metadata respectively. In the last step of the cycle, i.e. *processing*, Chapter 2 provides functionalities for query processing through its query mechanism.

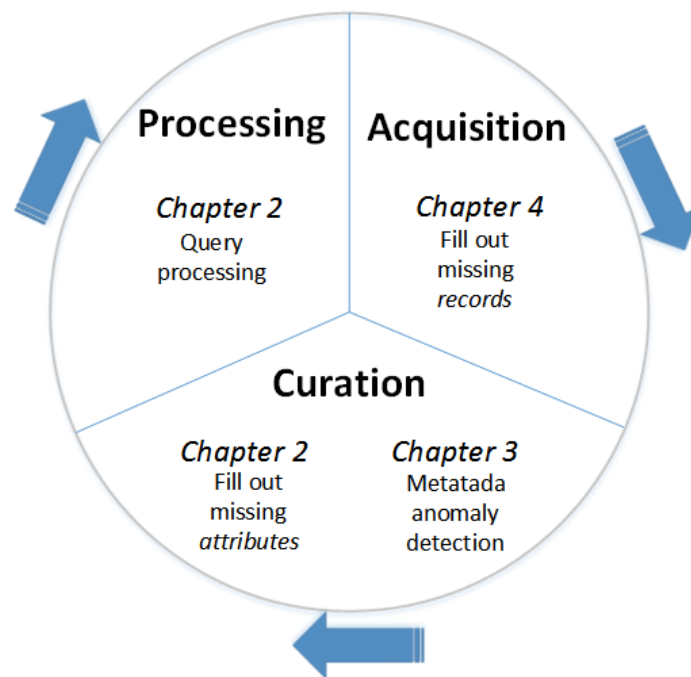


Figure 1.1: Overview of how chapters fit their roles in the Acquisition-Curation-Processing cycle.

This thesis was guided by some of the challenges faced by FNJV. Such challenges were identified in three main moments, as follows:

1. First, it was identified that many requests from scientists could not retrieve a desired information because of blank or null metadata fields even in expert-curated data. This scenario guided us towards improving metadata quality by detecting such gaps and filling out records with missing attributes. Moreover, our solution allowed extending metadata with additional environmental information, thereby letting biologists better understand the context in which an observation was made. This part of the research is reported in Chapter 2.
2. Next, still aiming at improving metadata quality, we focused on detecting anomalies in metadata. At this stage we provided an approach that performs spatial analysis by contrasting species observation sites against authoritative species distribution maps. This is reported in Chapter 3.
3. Finally, we identified obstacles related to lack of observation records fitting a scientist's requirements. Such requirements often change as scientists evolve their knowledge about some problem, creating a never ending need for fitting metadata requirements. Such scenario guided us towards developing an adaptive approach to acquire biological observations by using mobile devices, VGI and sensors. This is reported in Chapter 4

## 1.2 Organization and Contributions

Our work provides solutions to problems involving metadata quality. Our contributions were motivated by challenges faced by biologists on managing large amounts of animal sound recordings, using FNJV as a real life case study. The main contributions of this thesis are:

- Definition of a geographical approach for detection of anomalies in biological observation metadata;
- Specification of an architecture for retrieval of animal sound recordings. This architecture provides a query mechanism that fills out gaps in environmental context variables by using external data sources;
- Validation of the above by actual implementation of a prototype that runs over our web system to share biological observation metadata;
- Specification of a methodology to fill out gaps in biological observation databases by acquiring information from sensors and VGI provided by dynamic questionnaires;

- A framework that enables adaptive acquisition of VGI using mobile devices and sensors;

This thesis is organized as a collection of papers, as follows:

Chapter 2 corresponds to the paper “An architecture for retrieval of animal sound recordings based on context variables”, published in the journal *Concurrency and Computation: Practice and Experience* [34]. This chapter presents an architecture that provides a query mechanism to drive (a) a process of derivation of context variables and (b) process on metadata and derived contextual variables. In order to do this, the mechanism relies on multiple data sources, such as stored metadata, external semantic sources, external data sources and an ontology of vocalization context. The stored metadata represents the original collection. External semantic sources and external data sources represent third-party data collections that contain information to support the metadata derivation process (e.g., INMET [58] and DBPedia [9]). Context variables and their classification were constructed with help of domain experts. As a result, we constructed an ontology of context variables; (An ontology, here, is a data structure that organizes concepts hierarchically, and creates associations among them). The ontology of vocalization context is a local knowledge base that contains semantic rules to guide the derivation process.

The process of derivation of context variables fills out gaps in metadata records with missing environmental variable attributes. Queries perform information retrieval by considering both original and enriched metadata. Our case study with FNJV identified, for example, that only 7% of the metadata records contained air temperature information. Our approach allowed providing this information. As a consequence, biologists were able to correlate animal vocalizations and the temperature at the time the vocalizations were recorded – which is important to analyze behavior patterns.

Chapter 3 corresponds to the paper “A geographical approach for metadata quality improvement in biological observation databases”, published in the *Proceedings of the 9th IEEE International Conference on e-Science* [32]. This chapter presents an approach to improve metadata quality through detection of anomalies (e.g., faulty observations). The solution was based on spatial analysis that contrasted species observation sites against authoritative species distribution maps.

Our approach is composed by three steps: (i) preprocessing, (ii) finding anomalous places and (iii) presenting spatial output to the experts. In (i), we prepare the metadata for the spatial analysis by adding geographical coordinates in records that lack such information (based on both complete and incomplete textual location metadata). In (ii), we contrast the species observation locations against the species distribution maps. We also enable scientists to define a buffer area for the species distribution maps in order to overcome overestimation/underestimation problems in the expected species habitats. Finally, in step (iii) we present the results to the experts by showing maps with the *out-*

*liers* detected. Our strategy explores implicit knowledge provided by such outliers. We identified that such outliers are a starting point towards detection of anomalies in metadata. Biologists defined that such outliers can, in fact, indicate the following anomalies: A) metadata error; B) outdated metadata; C) errors in the species distribution range maps and D) possible new species pattern. In our case study with FNJV, experiments identified geographic anomalies for 12% of 1037 distinct species analyzed, with a total of 371 records out of 7049 records.

Chapter 4 corresponds to the paper “Adaptive acquisition of VGI to fill out gaps in biological observation metadata”, submitted to Proceedings of the 22nd ACM International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL) [31]. This paper, currently under review, is geared towards filling out gaps in the metadata. Here, a gap represents insufficient number of metadata records that fit a required scenario. Such gaps may occur, for instance, when a team of scientists wants to reuse data collected by another team, but the teams have distinct sampling methodologies. Gaps may also occur in the metadata as new information is acquired or as researchers revise their knowledge about some problem. Such dynamicity makes this a challenging problem.

Our approach can be divided in three steps: (i) detecting gaps in the metadata, (ii) acquiring VGI, and (iii) iterating the process. In (i) scientists define a suite of algorithms for gap detection. Such a suite detects gaps by considering the most important attributes in biological observation metadata, i.e., *what* (the species observed), *where* (location of the observation), *when* (date/time of the observation) and *how* (the methodology used). In (ii), scientists analyze the gaps provided by step (i) in order to define which kinds of VGI should be acquired. VGI is acquired after scientists define (a) a spatial area of interest for VGI acquisition (*where*); (b) the focus of the VGI (i.e., the target species, *what*) and (c) extra requirements (e.g., observations must be performed in a given time of the day - *when* and/or *how*). In (iii), scientists iterate the process, since we identified that metadata requirements might change as researchers acquire new knowledge about their problems.

Our methodology enables filling out metadata gaps by using information provided by VGI and sensors. This was implemented in a framework to create questionnaires to collect information from VGI. Such questionnaires can be changed on the fly as metadata requirements change.

Chapter 5 contains conclusions and some directions for future work.

Besides the papers in Chapters 2, 3 and 4, others were also published in the course of this thesis. Some publications are directly related to this research, while others supported additional research efforts. There follows a list of papers published, including the ones that compose this thesis.

- D. C. Cugler, C. B. Medeiros, and L. F. Toledo. Managing animal sounds - some

challenges and research directions. In *Proceedings of the 5th Brazilian eScience Workshop - 31st Brazilian Computer Society Conference*, July 2011.

- 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*, 25(16):2310–2326, June 2012 – (*extends the previous paper*)
- D. C. Cugler, C. B. Medeiros, S. Shekhar and L. F. Toledo. “A Geographical Approach for Metadata Quality Improvement in Biological Observation Databases.” In *Proceedings of the 9th IEEE International Conference on e-Science*, pages 212–220, 2013.
- D. C. Cugler and C. B. Medeiros. Adaptive Acquisition of VGI to Fill out Gaps in Biological Observation Metadata. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances In Geographical Information Systems*, ACM, 2014 (*under review*).
- D. C. Cugler, D. Oliver, M. E. Evans, S. Shekhar, and C. B. Medeiros. Spatial Big Data: Platforms, Analytics and Science (*under review*). Springer Intl. Journal for Spatially Integrated Social Sciences and Humanities (GeoJournal). Special issue on theory and spatial big data.
- M. S. Mota, J. S. C. Longo, D. C. Cugler, and C. B. Medeiros. Using linked data to extract geo-knowledge. In *XII Brazilian Symposium on GeoInfomatics (GeoInfo)*, pp. 111–116, Nov 2011.
- S. Shekhar, M. R. Evans, V. Gunturi, K. Yang, and D. C. Cugler. Benchmarking spatial big data. *Lecture Notes in Computer Science* 8163, pp. 81–93, 2014.
- R. B. Sousa, D. C. Cugler, J. E. G. Malaverri and C. B. Medeiros. A provenance-based approach to manage long term preservation of scientific data. In *30th IEEE International Conference on Data Engineering. Workshop on Long Term Preservation for Big Scientific Data*, pp. 126–133, 2014.





## Chapter 2

# An Architecture for Retrieval of Animal Sound Recordings Based on Context Variables

### 2.1 Introduction

Biodiversity research requires collecting several kinds of data, at multiple spatial and temporal scales, such as images, maps, sounds, texts and so on. This paper is concerned with recordings of animal sounds, in particular those recorded directly by experts (and not by sensors). In most cases, such recordings have metadata describing not only features from the sampled individuals, but also features of the environment where the sound was recorded. Sound files and metadata are managed in distinct ways.

Earlier animal recordings were commonly stored in magnetic tapes, requiring special attention to be kept clean, free of humidity and fungus infection. More recently, recordings use devices that save data in a variety of digital formats, such as ATRAC, AIFF, WAV and MP3, contributing to the heterogeneity problem [28]. Once the recordings have been stored in a repository in digital format, there begins a new series of data processing and management processes, and associated challenges.

Our work is motivated by the challenges faced by the Fonoteca Neotropical Jacques Vieliard (FNJV) at the University of Campinas (UNICAMP) [33]. Here, researchers have amassed the largest collection of animal sound recordings in the Neotropics<sup>1</sup>. All the records are offline, and only recently has there been a concentrated effort in creating digital metadata files for the recordings and converting them to digital media [39, 33].

In particular, this work involves research on sound databases, aiming to provide sup-

---

<sup>1</sup>The Neotropical region is one of the six biggest biogeographic areas in the world, defined based on its animal life features. It extends from the Southern Mexico Desert into South America [1].

port to retrieval of animal sounds based on context analysis. Challenges include the definition of appropriate context variables, the derivation of such variables from metadata, query processing and formulation based on these variables.

We are interested in sound management research aspects (as opposed to engineering aspects) from a database view point. Sound management systems are not new - e.g. see Motorola/Shazam Music Recognition Software [79]. What is then so different about the management of animal sound recordings? First, they are often made under difficult conditions, presenting lots of background noise. Also, when recorded in their natural habitats, many animal sounds appear in a single recording, there being a need to identify individuals (or at least individual species). Even if one assumes that recordings are performed under ideal conditions, vocalizations are very much sensitive to a wide range of contextual variables - e.g., time of the year, geographic distribution or even social aspects [54] [91].

Based on this scenario, there are several computational problems that need to be addressed to support biologists on retrieving animal sounds from a recording repository. One approach found in the literature is retrieval based on the analysis of acoustic features - e.g., by exploiting the physical properties of sound waves [77, 11]. However, acoustic properties of animal sounds vary widely, hampering this kind of retrieval.

Another way to retrieve sounds is to use explicit facts present in the recording metadata, such as information about species taxonomy, gender, age, location where the sound was recorded, time and date, and so on. Queries on metadata are limited to the stored fields, which are often incomplete or blank. Moreover, there is additional relevant information that is not explicit in the recording metadata and that is part of the context in which the sound was recorded - concerning engineering and biological/environmental variables.

Our work is geared towards supporting metadata-based retrieval. However, rather than limiting ourselves to recorded metadata, we exploit two new directions, that require tackling context and semantic issues:

1. Derive the contents of metadata fields that contain context information (e.g., environmental factors);
2. Derive additional context variables, not contemplated by the stored metadata, to enhance the scope of queries that can be supported.

The notion of “*context*” has many definitions (e.g. [49]), and several factors may influence the recording context. On the one hand, there are parameters involving the recording activity itself. For example, the quality of a recording may vary if this activity is performed by a specialist or by a novice. The methodology adopted to perform the

recording can also be part of the context. For example, the distance to the recording device may cause the animal call to be attenuated or degraded; or the quality of different devices may create noise in the recordings and capture undesirable environmental noise. These are examples of factors that may bias or cause errors in biological studies.

On the other hand, there are contextual variables that concern species and environmental conditions, and may influence the sound produced by the animals, modifying acoustic analysis. For example, male frogs call with higher repetition rate in the presence of females, when more males attend the breeding chorus, or in higher temperatures. Furthermore, vocalizations are sensitive to season, rainfall and time of day.

In our research the *context* of a recording considers only this second category of factors, i.e., those that concern species and environment. In other words, we do not concern ourselves with acoustic processing. This will require, among other things, to identify and characterize the context variables to be considered, as discussed later on in this text.

Summing up, the main contributions lie on designing and implementing solutions for retrieval of recordings of animal sounds from a recording repository. The main novelty, from the user’s point of view, is the support to queries that involve contextual factors. As discussed in subsequent sections, this requires research in, among others, query processing, ontology engineering and knowledge management.

The rest of this paper is organized as follows. Section 2.2 discusses related work. Section 2.3 introduces sound retrieval based on context analysis. Section 2.4 describes our current prototype using as case study the recordings of animal sounds from Fonoteca Neotropical “Jacques Viellard”. Section 2.5 presents our architecture. Finally, section 3.6 presents final considerations and ongoing work.

## 2.2 Related Work and Some Challenges

Figure 2.1 shows the basic flow of processes concerning scientific data management, from collection to analysis. They need to be detailed for each domain.

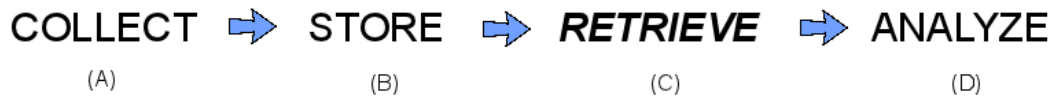


Figure 2.1: Steps involved between collect and analysis processes. This research concerns step (C).

Firstly, in our domain, biologists record animal sounds and metadata related to the collection step (Figure 2.1, A) - e.g., data about climate conditions, site, scientific name of the animal, and so on. On the next step (B), the collected information is archived

(most often, metadata are stored in spreadsheets and sound recordings in individual files). Thereafter, biologists access the information in order to retrieve the desired recordings (C), usually by going through metadata, and perform some kind of analysis over the retrieved data (D).

Sound archival is just one of the problems that surround research on sounds [46]. Managing such data in biodiversity studies, in an integrated way, poses very many challenges [33], ranging from conceptual to physical levels and requiring all kinds of algorithms and tools, which are typical of e-Science environments [53]. Proposals to handle such requirements have to consider an entire computational environment, in which semantics play an important role – e.g., [42].

The analysis of related work shows that the use of ontologies may add semantic information to the recording metadata. Ontologies can therefore be a key aspect to support animal sound retrieval based on analysis of the context.

### 2.2.1 Sound Retrieval Based on Recording Metadata

A considerable amount of animal sound metadata provided by biologists are incomplete, hampering queries. This is particularly true for old recordings, where spatial information is often imprecise. With the advent of GPS, the situation is improved, but even then lots of metadata depend on manual insertion during recording expeditions (e.g., information on environmental variables, such as temperature and rainfall). Identification of the species being recorded is also a hard task, since many times the scientists do not see the animals whose vocalization they are recording. For this reason, metadata may, for instance, lack an animal's scientific name, providing only partial taxonomic information. This all adds up to gaps in the metadata, compounded by the difficulty of manually registering relevant facts.

In spite of the problems cited above, the simplest solution adopted by scientists is to retrieve recordings based on metadata values. This leads to severe problems when related to missing information. Sometimes queries cannot retrieve a desired information because a filter parameter points to a metadata field that is blank or null. In this case, the queries may retrieve false-positive records, ignoring ones on situations where they can be relevant for the search space.

Another problem is created when recording metadata is saved in spreadsheets. Although biologists commonly use spreadsheets as a data repository, there are not enough tools or resources that enable a reliable and consistent representation of the data – e.g., to represent relationships among fields. Using spreadsheets with this purpose limits the scope of metadata management, hampering the execution of more complex queries.

### 2.2.2 Sound Retrieval Based on Acoustic Similarity

A considerable amount of work on sound retrieval either focuses on techniques of acoustic similarity, or uses acoustic similarity to support new retrieval techniques, using specific algorithms for different kinds of sounds. Sounds are waves, hence sound retrieval based on acoustic similarity analyzes similarity among different sound waves. Figure 2.2 exemplifies this concept, showing spectrograms of four different call types of the same treefrog species.

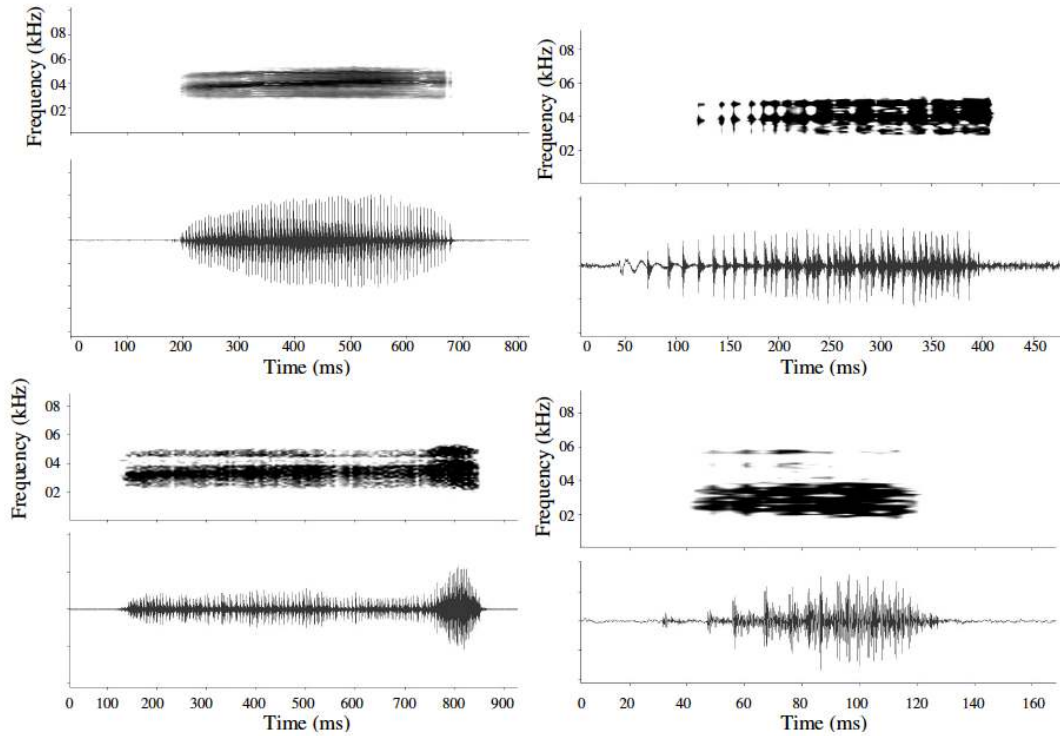


Figure 2.2: Four different call types of the same treefrog species (*Scinax fuscomarginatus*), emitted in different social contexts (modified from [90]).

There are many kinds of input provided in such retrieval requests. For instance, Kotsifakos et al. [64] created an algorithm based on acoustic analysis to retrieve music from a database through queries based on music hummed by users. There is also work that uses acoustic analysis for speech recognition. For instance, Wang et al. [93] use, among other technologies, MFC coefficients (*Mel-Frequency Cepstral*) to perform acoustic analysis of sounds produced by humans, in order to automate residences using voice commands. In a different domain [87], there is the description of a system that analyzes vocalizations produced by pigs for monitoring of stress. The system can identify different feelings (e.g., anger) based on acoustic analysis of vocalizations.

In devising patterns for animal sounds, many issues have to be considered. Due to the complexity of such sounds, retrieval techniques based on acoustic similarity alone may not be effective - queries need to consider additional information or specific algorithms may need to be developed for distinct taxonomic families of animal groups<sup>2</sup>. Examples of the need for specific algorithms are:

- crickets, fishes and frogs generally present simple innate calls (i.e., just one type of note per call type) with well defined structure (even between individuals of different populations);
- birds and mammals may present different notes in a single call (complex calls), that may be subject to learning;
- habitat fragmentation may influence the frequencies of animal vocalizations [43, 94].

With the increase of computational power of personal computers, many softwares for acoustic analysis were developed. One example of recent research is the work of Wimmer et al. [97], a workbench that, among other functionalities, supports query and recognition of animal sounds using acoustic similarity (based on ASR - Automatic Speech Recognition). In another research, Bardeli [11] created a method for acoustic similarity search in animal sound databases, also proposing algorithms for feature extraction, indexing and retrieval for these databases. Similarly, Gunasekaran and Revathy [47] proposed an algorithm for classification and retrieval of animal sounds. The algorithm is based on fractal dimensions analysis to compute the k-nearest neighbours.

Automatic recognition is not generic and has limitations – e.g., algorithms for specific animal groups may not cover important features that are essential for other groups. Hence, there is a lack of generic systems covering the characteristics of different groups. Such systems would allow the execution of many types of queries required by biologists, which are not currently supported.

### 2.2.3 Sound Retrieval Based on Semantic Similarity

Sound retrieval based on semantic similarity appeared in order to overcome some shortcomings present on retrieval based on acoustic similarity. It allows representing explicit and implicit information associated with animal sounds. The use of ontologies is the most common means of adding semantics to sound metadata, allowing the execution of queries that cannot be performed using just acoustic similarity. There are many research efforts

---

<sup>2</sup>The term “group” is frequently used by biologists to identify/classify sets of species according to specific features. Therefore, in this paper, an “animal group” is not a random group of species or individuals.

concerning semantic similarity for several domains. We emphasize that we have not found research for semantic retrieval of animal sounds.

The system described in [13] retrieves generic sound records - such as exterior sounds, water, cars, agricultural machinery, horses, dogs, schools - ranking items in a database using semantic similarity. Ontologies are used as part of query processing to narrow the search space - strategies, algorithms, mechanisms. In order to improve the results, queries are performed using both acoustic and semantic metadata queries.

Also focusing on semantic similarity, Mechtley et al. [76] proposed to use words from WordNet [37] to allow users to tag sounds. They incorporate these tags to find semantic similarity among the concepts used to describe sounds, and proposed a network structure for integrating similarity between the semantic tags. Then, they use acoustic similarity algorithms and semantic tag similarity to retrieve similar sounds from a query-by-example system - the example can be a sound or a tag.

Buchanan’s research also focuses on semantics [24]. It considers the task of attaching meaningful representations to recordings of different kinds of sounds (such as trains, babies crying, dogs, rain). For example, for the query “*lion roar*”, the system should retrieve records related to lions roaring. The goal is to show that this approach to model the semantics of sound is appropriate for the task of classifying and retrieving audio.

All research cited here uses semantics to improve sound retrieval for several domains. However, as already mentioned many times in this paper, animal vocalizations have specific features (related to the context in which they were recorded) that need to be taken into account and that are not faced by them.

## 2.3 Defining Context Variables to Support Sound Retrieval

Information related to the *context* in which the sound was recorded is not considered in retrieval based on acoustic or semantic similarity. We were not able to find any report that specifically considers context in retrieval. At most, contextual variables are stored in metadata fields. In this case, retrieval is based on metadata values. Related work might then consider some sort of semantic retrieval (section 2.2.3), when metadata include links to domain ontologies. Another kind of related work involves expanding queries on metadata by using taxonomic ontologies [74]. Finally, our work on this kind of metadata-based retrieval and the associated prototype (see section 2.4) are yet another instance of this research direction.

However, none of these retrieval modalities considers context information. Context analysis, in the sense used in this paper, involves several factors related to the animal

and the environment in which the sound was recorded. Designing an ontology of context variables for sound management is one of the challenges that we face. So far, we have been able to identify three major classes of context factors that this ontology should include - temporal, social and environmental - exemplified next.

Besides examples given previously, context variables should be considered to treat situations such as:

- Temporal factors
  - hour after/before sundown: the daily activity of several species is regulated by the sunrise and sundown; individuals will call before, during or after sunrise/-sundown, depending on the species.
  - lunar phase: nocturnal species may present conspicuous behaviours – i.e., more calling activity – when the moon is new, and secretive behaviours – i.e., less calling activity – when the moon is full.
- Social factors
  - isolated/grouped individuals: when isolated, individuals may call more sporadically.
  - distance between individuals: as the distance decreases, the power of the call decreases and repetition rates increase.
  - mating season: some species only emit calls (or produce more calls) during the breeding season, in comparison to non-breeding season; it is common to observe that in the beginning of the breeding season the calling activity is low, reaching the peak soon after and then declining slowly throughout the remaining breeding period.
  - presence/absence of females: males will mostly produce courtship calls only in the presence of females.
- Environmental factors
  - earthquakes: species can vocalize just after, or during earthquakes to inform other individuals about the imminent danger.
  - wind speed: for species that may lose body water, windy conditions may cease or reduce the calling activity.
  - rainfall: rainfall generally stimulates calling activity of several species, but when the rain is severe, it may cause the individuals to be silent.



- temperature: the higher the temperature, the higher the calling activity of several species.
- habitat fragmentation: the higher the fragmentation of the habitat, the more the individuals are exposed to open fields conditions, which may influence the frequencies produced or recorded.

Moreover, the *geographic region* is an important contextual information that influences environmental factors. Thus, this information should always be taken into account when processing context variables.

We point out that many such concepts are directly defined in consensual ontologies – e.g., NASA’s SWEET, on phenomena, or SEEK on ecological variables. As explained in Section 2.5, these are considered to be remote semantic sources that will also be involved in the derivation of contextual information and expansion of user queries.

Except for the geographic region, almost none of these facts are recorded in sound metadata. At most, some of these facts are provided via free text. Temperature is sometimes included, but even then, not on a regular basis – e.g. in our case study, only about 7 percent of the metadata records contain temperature information. As will be seen, our work tries to derive missing information by taking advantage of combining concepts in the ontology with accessing external data and knowledge bases, as well as of initiatives associated with the Linked Data Paradigm. The notion of Linked Data appeared in the Semantic Web context. The term is related to a set of practices for publishing and sharing structured data on the Web. Basically, Linked Data uses the RDF (Resource Descriptor Framework) format to make typed statements that link things [17, 18]. The four “rules” of linked data are: (i) Use URIs as names of things; (ii) use HTTP URIs so that people can look up those names; (iii) when someone looks up an URI, provide useful information; and (iv) include links to other URIs, so they can discover more things [69]. Perhaps the best example of a Linked Data system is DBpedia, which provides Wikipedia information in a structured way [9].

Section 2.4 shows our preliminary work on metadata cleaning and metadata-based retrieval. Section 2.5 presents our architecture for a context-based sound recording retrieval system.

## 2.4 Publishing Animal Sounds on the Web - Case study and prototype

Since 1970 researchers from FNJV - Fonoteca Neotropical “Jacques Vielliard” - UNICAMP - recorded about 40.000 animal vocalizations [39], mainly birds and amphibians

in South America, mostly recorded in magnetic tapes. These recordings are undergoing digital conversion. Although this conversion started years ago, due to its complexity, about 15.000 recordings from the collection have been saved in digital format so far. This collection - the seventh largest in the world - frequently receives visits from biologists for their research, and is being progressively augmented by donations.

Currently, spreadsheets are used to manage the metadata of FNJV recordings and there is no full-fledged system to manage recordings and metadata. These facilities are not adequate for more advanced management needs. Another problem is the fact that the collection does not support online access, which hampers its widespread use. Thus, our first step was to develop a web system (available at <http://proj.lis.ic.unicamp.br/fnjv>) to provide basic functionalities for animal sound management, such as: to publish metadata on the web; to allow scientists to upload and download recordings; and to retrieve sound records through textual search based on metadata. The activities needed to develop and publish these data cover the four steps of Figure 2.1.

Part of the data storage stage (Figure 2.1 - Step B) concerned “syntactic” data cleaning, checking for basic inconsistencies in the metadata. For example, there were different patterns for date – e.g. 22/11/2011, 22/11/11, 11/22/2011, abbreviations of place names and others.

The prototype was developed in Java and JSF (Java Server Faces). Other technologies were also used: Ajax (Rich Faces Framework) and the PostgreSQL DBMS. This DBMS was used in order to support integration with other tools developed at our laboratory [3].

Currently, the prototype can be accessed using Firefox 3.6 or Internet Explorer 9 (or later). It has the following functionalities:

- Public functionalities
  - Navigate the whole collection of sounds
  - Perform metadata-based queries using taxonomic parameters
- Private functionalities
  - Upload sounds
  - Select specific records and export into spreadsheets
  - User authentication and management operations

### 2.4.1 Public functionalities

#### Navigate the whole collection of animal sounds

This functionality allows users to see all recordings, ordered by scientific taxonomy. Though simple, this kind of functionality is a first step towards a more sophisticated nav-

igation system, whose specification requires user feedback. Figure 2.3 exhibits a screen copy showing how biologists are able to view scientific taxonomy data associated with each recording. For instance, record with FNJV identification 7 concerns vocalization of a bird: Phylum *Chordata*, Class *Aves*, Order *Tinamiformes*, Family *Tinamidae*, Genus *Tinamus* and Species *tao*.

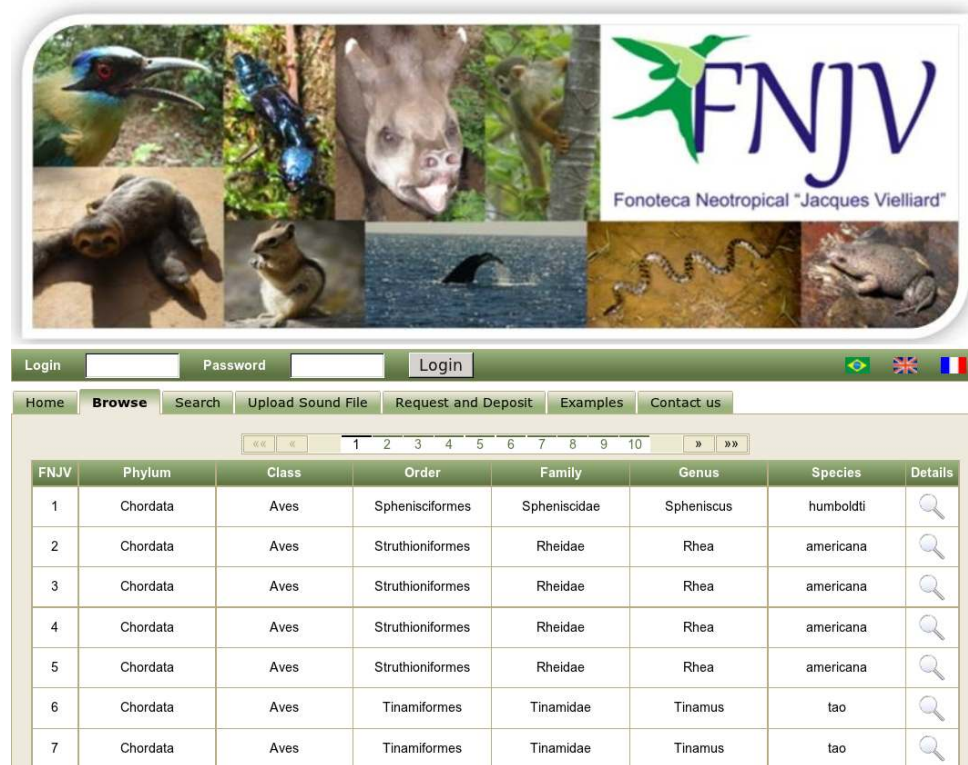
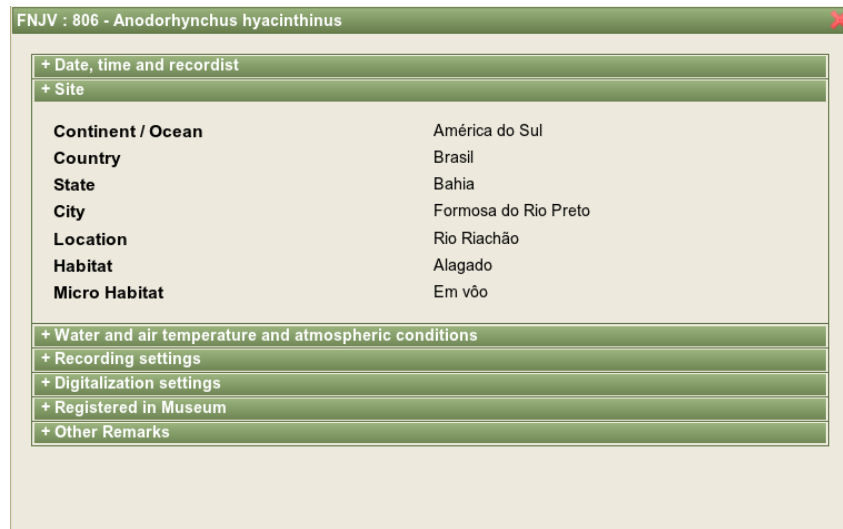


Figure 2.3: Partial screenshot of the page that allows any user to navigate through the whole collection of animal sounds.

Users can also view all recording metadata by clicking on the “details” button. Figure 2.4 shows the kinds of metadata of a selected sound record - e.g., information such as the time and date in which the sound was recorded, country, state and city, temperature of the air at the moment of the recording, and so on. In particular, the figure shows information on metadata for the site where the recording took place.

### Metadata-based textual queries

This functionality allows biologists to retrieve only the desired sound records. Figure 2.5 shows the query form, with textual filters. Currently, the filter is restricted to scientific taxonomic terms. For example, Figure 2.5 shows a query that retrieves all three recordings from animals in the Phylum *Chordata*, Class *Aves*, Order *Tinamiformes*, Family



FNJV : 806 - Anodorhynchus hyacinthinus	
+ Date, time and recordist	
+ Site	
Continent / Ocean	América do Sul
Country	Brasil
State	Bahia
City	Formosa do Rio Preto
Location	Rio Riachão
Habitat	Alagado
Micro Habitat	Em vôo
+ Water and air temperature and atmospheric conditions	
+ Recording settings	
+ Digitalization settings	
+ Registered in Museum	
+ Other Remarks	

Figure 2.4: Details about metadata of one recording - geographic metadata on recording site.

*Tinamidae*, Genus *Tinamus* and Species *guttatus*. Users can subsequently retrieve more information (Figure 2.4).



Home	Browse	Search	Upload Sound File	Request and Deposit	Examples	Contact us	
Phylum <input type="text" value="Chordata"/> Class <input type="text" value="Aves"/> Order <input type="text" value="Tinamiformes"/> Family <input type="text" value="Tinamidae"/> Genus <input type="text" value="Tinamus"/> Species <input type="text" value="guttatus"/>							
FNJV	Phylum	Class	Order	Family	Genus	Species	Details
3991	Chordata	Aves	Tinamiformes	Tinamidae	Tinamus	guttatus	
10478	Chordata	Aves	Tinamiformes	Tinamidae	Tinamus	guttatus	
11662	Chordata	Aves	Tinamiformes	Tinamidae	Tinamus	guttatus	

Figure 2.5: Form to perform metadata-based queries

## 2.4.2 Private functionalities

Some of the current functionalities of our prototype are available only for authorized researchers. This approach was adopted given confidentiality and privacy issues. In the long run, some of these functionalities should become available for all users.

## Upload sounds

This functionality allows authorized users to upload sounds. Expert curators must mediate the deposit of recordings. This may eventually be used to support *citizen science* [59], e.g. people who like to record animal sounds as a hobby will also be able to publish and share their recordings, and help scientists in their work. In this case, the recordings will be double-checked by experts prior to their deposit in the collection.

## Select specific records and export into spreadsheets

Researchers may want to select some specific information to use in their research. We thus provided a functionality that allows them to select desired metadata information and then export these metadata into spreadsheets. Figure 2.6 shows the selection of two records and the link to export them into spreadsheets. This kind of format was chosen due to its widespread use by biologists.



FNJV	Phylum	Class	Order	Family	Genus	Species	Details	Select
8636	Arthropoda	Arachnida	Araneae	Theraphosidae	Acanthoscurria	gomesiana		<input checked="" type="checkbox"/>
8637	Arthropoda	Arachnida	Araneae	Theraphosidae	Acanthoscurria	gomesiana		<input checked="" type="checkbox"/>

Figure 2.6: Page that shows the records selected by the user and the link to export them into spreadsheets

## User management facilities

This considers a set of functions that allow authorization of users, password setting, and management of users' roles.

### 2.4.3 Rationale behind the prototype

We have adopted a collaboration policy that has proved to be successful in previous multidisciplinary projects of ours<sup>3</sup>. Instead of designing and developing a complex system with many query possibilities (e.g., sound mining, or indexing), we developed a first prototype to help scientists manage their data. Once they test and approve a prototype, then we can improve on it, on a continuous evolution lifecycle, with rapid prototyping. This has the advantage of ensuring fast feedback and in attracting a larger number of users. We point out that, even if relatively simple, this kind of web system is very

<sup>3</sup><http://www.lis.ic.unicamp.br/projects>

much in demand by experts and, in our case, it has the advantage of the richness of the underlying data collection.

This first prototype also enabled us to familiarize ourselves with the target domain. Besides that, it provided an environment where we can query cleaned sound metadata, freeing us from direct manipulation of spreadsheets.

## 2.5 Semantic retrieval of animal sounds based on context analysis

This section presents our architecture and the present stage of its implementation. We start by an overview of the architecture, present details via an implementation of some modules, and finish with examples of query processing.

### 2.5.1 An architecture for retrieval of animal sounds based on context analysis

Figure 2.7 gives an overview of our architecture. Though centered on processing of animal sounds, it can also be adapted for other domains. The figure shows that there are two types of data sources: internal and remote. Remote sources are accessed via a communication layer (item 12) that translates access requests from our query mechanism into specific requests to the different sources. In other words, it provides distinct communication mechanisms to linked data sources such as DBPedia and to standard sources that are accessed via web services. Internal sources comprise the sound repository, sound metadata and the ontology of vocalization context.

#### User modes

The “query mechanism”, Figure 2.7 - box (3), serves a dual role - (a) it fills in missing metadata, and (b) it processes user queries. Both roles are executed analogously, except that for role (a) data curators invoke the mechanism directly, and for role (b) scientists in general (including curators) request retrieval of sound recordings based on metadata and context variables.

(a) Curator – The curator prompts the query mechanism to derive missing contextual metadata fields (item 9), indicating which parameters are to be filled for all records (e.g. temperature, country, state). Each record is returned to the curator, who either accepts the system’s suggestion or not (item 10). If accepted, they are stored.

(b) Scientist – The scientist starts the query process (item 1) specifying desired options for the query filter. These options can be both stored metadata fields (e.g., scientific name)

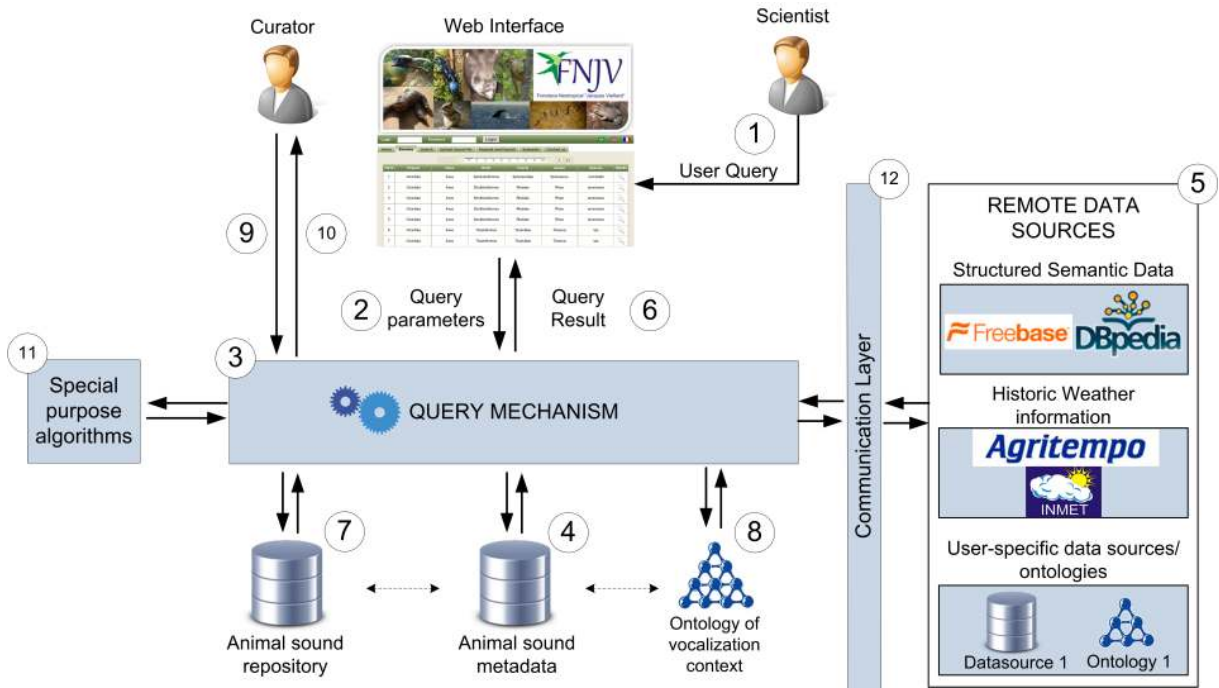


Figure 2.7: Semantic retrieval of animal sounds based on context analysis.

or derived contextual data (e.g., rainfall indexes or lunar phase).

Input parameters feed the query mechanism (2) and (3). After processing, the mechanism delivers the query result (6) to the user. The web interface presents metadata records with links to the respective sound files from the animal sound repository (7).

On both cases of user profile (a) and (b) we are not concerned with aspects of interface design or query flexibility, which escape the purpose of this research. Our approach is the following. For the curator mode, the labels of all contextual metadata fields are presented to the curator, who chooses the ones to be completed, if blank or unknown – e.g., temperature or country. The interface for the scientist mode allows scientists to fill in values for the following parameters: any metadata field (e.g., taxonomy, recording device) and any contextual variable that appears in the ontology of vocalization context (namely social, environmental, temporal and geographic region variables). This ontology, as explained next, will drive the process of derivation of context variables.

### The query mechanism - deriving contextual information

The role of the query mechanism is to meet user requests. It drives the process of derivation of context variables and processes queries on metadata and derived contextual variables.

In order to do this, the mechanism relies on the following data sources: stored meta-

data (4), external semantic sources (5), external data sources (5) and the ontology of vocalization context (8).

The derivation of context variables uses two complementary mechanisms: customized modules that embed special purpose algorithms (11) and ontology-based processing, in which rules associated to the ontology (8) drive the derivation process. For instance, a customized module can be used to compute an average windspeed factor for a region based on a set of readings of sensors, retrieved from an external source. An example of an ontology based derivation is when concept “ridge” is defined to be a feature associated with concept “relief” which itself can be found in the Freebase knowledge base [20, 40]. Still another example of ontology based derivation is a rule that says that the value of concept “temperature” can be derived from values of concepts “geographic location”, “date” and “time”, and invoking some kind of service.

The core of the derivation process concerns the interaction between the special purpose algorithms (11) and the ontology (8). The ontology of vocalization context has four main classes of concepts (see section 2.3): temporal, environmental, social and regional. It contains one entry for each context variable (a concept) that can influence vocalization, and is part of ongoing research. Its entries can belong to one of three types (the type being one of the properties defined for each variable):

- Type 1 - the value of that variable can only be computed by invocation of a special purpose algorithm. The address of this algorithm (11) is a property of the variable. The algorithm may need to access external sources;
- Type 2 - the value of that variable can be computed by applying ontology inference rules written in OWL + SWRL (possibly also involving access to external semantic sources, such as DBPedia or NASA’s SWEET ontology or SEEK ontology);
- Type 3 (hybrid) - the value of that variable is computed by first executing an inference rule (as in Type 2) and then use the result of the inference to invoke a special purpose algorithm (as in Type 1)

We now proceed to explain how the architecture blocks fit together. Let us first examine curator requests for filling missing metadata fields. For each such field, the query mechanism starts from the ontology to find out how to derive a value. It drives the interaction of ontology rules and special modules, according to the Type of a concept, to derive these values, which are then presented to the curator for validation. The curator must analyse and validate all retrieved information before it is saved in the animal sound metadata repository. In order to maintain the original metadata, derived information must be saved separately.



The second part of the explanation is related to the queries performed by scientists (1). At this step, the scientist must define the query parameters (2). After the scientist selects all parameters to compose the query filter, the query mechanism:

(a) Queries the animal sound metadata (4) using SQL. It thus narrows the search space by selecting records which meet predicates on metadata, resulting in a subset of metadata records  $Q$ . Obviously, if  $Q$  is empty, the query cannot proceed and an empty result is returned to the scientist.

(b) If  $Q$  is not empty, then for each record  $q$  in  $Q$  the query mechanism drives the interaction of ontology rules and special purpose algorithms described previously, using stored metadata values, when applicable, to derive additional contextual information associated with  $q$ .

(c) The final result returned to the scientist is a subset of records  $Q' \subseteq Q$  in which the additional contextual information matches input filters.

We next present a prototype we developed to illustrate the feasibility of this approach.

### 2.5.2 Extracting information from remote sources to derive context variables

This section describes the present state of our implementation. It is based on invoking the special purpose algorithms of the architecture (Figure 2.7 - item 11). This implementation does not include any context ontology in the processing, which is left to future work.

Table 2.1 shows 27 (out of 51) metadata fields that are present in the FNJV collection, with one example of a record that describes information related to the vocalization of a frog. We notice, for instance, that some fields have precise information while others are left blank (e.g., atmospheric conditions) and others are indicated as “Ignored”. The semantics of “Ignored” and blank sometimes are the same (i.e., not entered by the expert) or may mean different things.

The architecture of the prototype is presented in Figure 2.8. Here, the process to fill metadata fields and derive context variables starts from gathering information about the location where a sound was recorded<sup>4</sup>. Once we know when and where the sound was recorded, we can find many contextual facts. In the example cited in Table 2.1, the location where the sound was recorded corresponds to the city of Cabreúva, São Paulo state, in Brazil. It is retrieved from the animal sound metadata (Figure 2.7 - item 4 and Figure 2.8 - item 1). This information is used to generate a query in Metaweb Query Language (MQL) [19] (item 2), in order to query the Freebase knowledge base [20, 40]

---

<sup>4</sup>Here, a general ontological rule for vocalization context would say that the tuple <location, collection time, collection date> will provide information on several contextual variables. Another rule will state that location (in latitude/longitude) should be obtained from <City, State, Country> and so on.

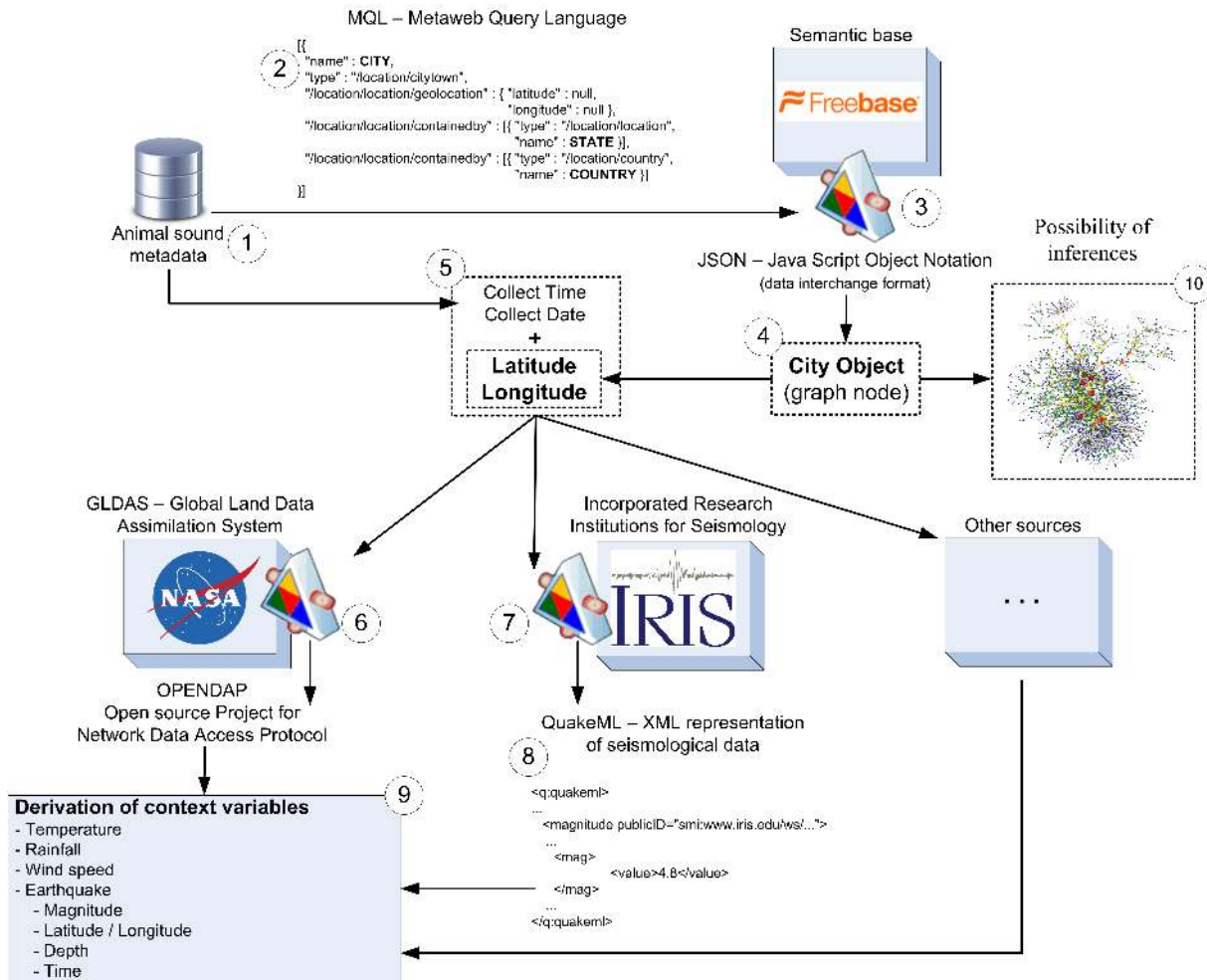


Figure 2.8: Architecture of prototype to derive contextual information - use of special purpose algorithms, external knowledge bases and data sources.

(item 3). All Freebase queries are answered in Java Script Object Notation (JSON) [61] data interchange format. Freebase provides the answer as a graph node (item 4) of Freebase’s knowledge base - a city object in Freebase vocabulary.

The next step is to collect latitude and longitude of the location represented by the node. Freebase represents a location as the concept “/location/citytown” and latitude/longitude as the concept “/location/location/geolocation”.

Once latitude/longitude are obtained, they are combined with stored metadata fields “collect time” and “collect date”, thus providing the <where, when> used to derive missing metadata and new context variables. “Where” and “when” are then provided as input to web services available on the web, from which such variables can be extracted. The figure shows two such services (NASA’s GLDAS and IRIS) which are part of our implementation.

Additional sources can be also added (e.g., Brazilian databases of historical meteorological time series).

Figure 2.8 also shows that the graph node retrieved can be used as input for further inferences in Freebase’s graph – e.g., to find the relief on a region.

In more detail, NASA Global Land Data Assimilation System (GLDAS) [84] provides plenty of information derived from satellite images from 1979 onwards – e.g., soil moisture, surface pressure. The figure shows that given this input (5), GLDAS can provide temperature, rainfall and windspeed information with one degree spatial resolution and three hours temporal resolution. GLDAS service can be accessed through HTTP, and queries return information in Open Source Project for Network Data Access Protocol (OPENDAP) format [30].


IRIS stands for Incorporated Research Institutions of Seismology [2] (item 7). Seismological information can be retrieved on IRIS web service through a Java API. Information such as earthquake magnitude, latitude and longitude of the epicentre, time and date and depth of the earthquake is represented in QuakeML format (XML representation of seismological data) [85] - item 8.

Figure 2.9 shows a screen copy of this prototype, developed to simulate the process shown in Figure 2.8. All information shown in this figure is related to the example cited in Table 2.1, which corresponds to the city of Cabreúva and collect date 08/16/1986. To run the prototype, firstly it is necessary to fill in information about the place where the sound was recorded. After that, clicking on “Query Freebase” the prototype queries the Freebase knowledge base and retrieves latitude and longitude for the location (if available in Freebase), filling latitude/longitude fields on the right upper part of the figure. Alternatively, one can provide the latitude/longitude directly. If, moreover, date and time are provided (other parameters at top right), the user must click on the button “Query NASA and IRIS”. After some seconds, the prototype retrieves from NASA and IRIS services the air temperature, rainfall volume, wind speed and earthquake information for the latitude/longitude, time and date informed. Moon phase information is not derived from NASA and IRIS services. It is acquired through a mathematical function provided by a JAVA API (a special purpose algorithms module - Figure 2.7 - item 11). Finally, the location is shown in a map at the bottom.

We point out that none of this information is present in Table 2.1. Temperature can be now added into the metadata and, together with the other contextual variables, support new queries. We chose Freebase as a starting point because of the navigation and derivation possibilities it offers. Likewise, this first prototype invokes GLDAS and IRIS as good examples of reliable and open services. For example, DBpedia [18] could also be used in order to find environmental information associated with the location where the sound was recorded [78], through the use of the Linked Data paradigm.

Freebase service	NASA and IRIS services
City.....: <input type="text" value="Cabrêuva"/>	Latitude/Longitude...: <input type="text" value="-23.3075"/> / <input type="text" value="-47.1331"/>
State.....: <input type="text" value="São Paulo"/>	Date (dd/mm/yyyy): <input type="text" value="16"/> / <input type="text" value="08"/> / <input type="text" value="1986"/>
Country...: <input type="text" value="Brazil"/>	Time (from 0 to 7) - 0AM(0), 3AM(1), 6AM(2), 9AM(3), 12PM(4), 3PM(5), 6PM(6), 9PM(7): <input type="text" value="7"/>
<input type="button" value="Reset"/>	<input type="button" value="Query NASA and IRIS"/>
Status: <b>Service performed successfully</b>	

DERIVATION OF CONTEXT INFORMATION FROM RECORDING METADATA	
Air temperature (°C): <b>21.0</b>	Earthquake info: No data found for this location and period
Rainfall volume in 3 hours (mm): <b>0.018576</b>	
Rainfall daily volume (mm): <b>3.807</b>	
Instantaneous wind speed (m/s): <b>7.34</b>	
Moon phase: <b>Lua crescente convexa</b> 	


  


Figure 2.9: Prototype to query Freebase, NASA and IRIS services.

Nevertheless, for each such choice we had to implement dedicated modules, which hampers extensibility and scalability. The use of the ontology of vocalization contexts will provide such facilities. Nevertheless, there will always exist a need for customized modules, either because ontologies cannot solve everything, or even for performance reasons.

### 2.5.3 Extracting information from semantic bases in order to improve queries

Let us now examine a simple real example: *retrieve all sounds from *Rhinella ornata* toads recorded when the temperature was higher than 23°C*. Recall that in our case study only 7 percent of the records contain temperature information. In this case, curators must request to fill in metadata temperature fields from remote data sources – as shown in Section 4.4.3. Afterwards, any query can comprise temperature information on its parameter.

In our case study, the stored metadata have no moon phase field. Based on this scenario, another real example supported by our architecture is: *retrieve all sounds from *Rhinella ornata* toads recorded in a specific moon phase*. In this case, based on the

time and date of collect, the architecture computes the moon phase when the sound was recorded, and retrieves only *Rhinella Ornata* vocalizations recorded in the moon phase defined in the query. To the best of our knowledge, there is no system that supports such features.

## 2.6 Final Considerations and Ongoing Work

The work reported in this paper is motivated by challenges faced by biologists on managing large amounts of animal sound recordings, using FNJV as a real life case study. The current collection, which has more than 1,5 terabytes of digital sound data, is growing up. Only half the recordings have been digitalized. We are now working on publishing all metadata online, linked to the recordings.

Ongoing and future work involve attacking some of the challenges mentioned. First, we are continuing to investigate context variables, and rules/algorithms to derive them, to come up with a full-fledged context ontology. Its concepts will direct the final query interface.

Another problem to consider: additional issues concern finding out appropriate context data sources, and designing mechanisms to process query filters.

Though we provide functions to fill missing metadata from recording metadata, the values provided may be imprecise (e.g., temperature or rainfall concern a region in which a recording took place, but may not necessarily apply to a specific spot). Thus, query processing on missing metadata must inform the user of this fact. This, in turn, may need to consider aspects of query processing in probabilistic databases. We also intend to work on a specification to interoperate with other animal sound recording sites, via import/export of metadata in a consensual format - e.g., RDF.

One additional issue to be considered in the future is security. Many sounds are copyrighted, but can be made available for biologists who want to use them for research. To do this, it will be necessary to provide distinct access control mechanisms. This control may be also exercised for all downloads, and will create new data files that can be further used for data mining. For example, scientists can find other scientists that are working with the same species calls, improving interaction in the scientific community and improving worldwide science.

	METADATA FIELD NAME	EXAMPLE OF A RECORD
1	Phylum	Chordata
2	Class	Amphibia
3	Order	Anura
4	Family	Bufonidae
5	Genus	Rhinella
6	Species	ornata
7	Recordist	Gilda V. Andrade
8	Collect time	20:00
9	Collect date	08/16/1986
10	Country	Brazil
11	State	São Paulo
12	City	Cabreúva
13	Location	Cava Farm
14	Habitat	Ignored
15	Micro-Habitat	Ignored
16	Number of individuals	One
17	Kind of contact	Ignored
18	Gender	Undetermined
19	Kind of vocalization	Sound
20	Distance from the animal (mts)	-
21	Air temperature (°C)	-
22	Atmospheric conditions	-
23	Recording device	Uher 4000 Report I
24	Microphone model	M538
25	Captivity	No
26	Sound file format	WAV
27	Frequency (kHz)	48

Table 2.1: Subset of metadata fields that are present in the FNJV collection and an example of a metadata record.

## Chapter 3

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

### 3.1 Introduction

Our work is geared towards 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)” [23]. 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 the 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.

Many publications on curation of scientific metadata – e.g., [5, 99] – 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 [26, 82]. 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 hard because lots of metadata derive from the observation methodologies adopted. Such observations often result from many scientific expeditions undertaken along the years. As remarked by [23], for instance, since observation records depend on such teams, they suffer from both schema and semantic heterogeneity (i.e., structure and content). There is not only 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, 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., [99]), there is little concern with uncertain and imprecise descriptions of locations (e.g., via place 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 3.1. 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.

Our approach is evaluated using a case study at the Fonoteca Neotropical Jacques Vielliard (FNJV) [39], one of the top 10 animal sound collections in the world [83]. 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.



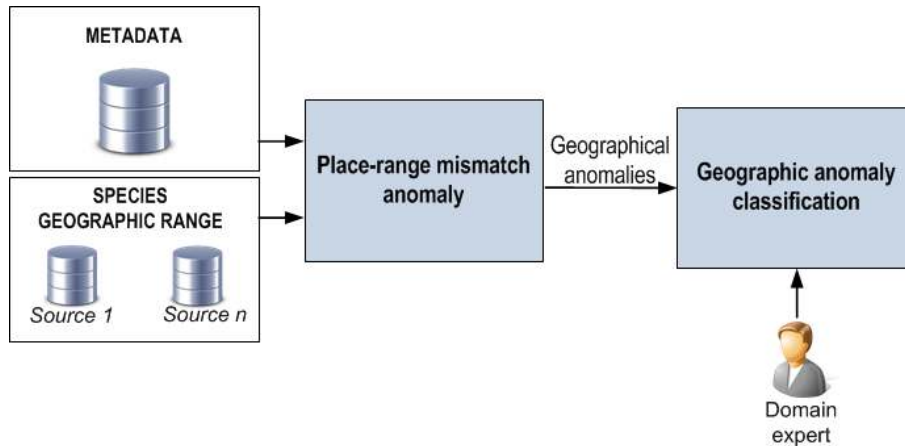


Figure 3.1: Overview of our geographical approach.

## 3.2 Background and Related Work

### 3.2.1 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 [38].

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 biological 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 [29] is an international center for the study, appreciation, and conservation of birds [21]. Fonozoo [88], Fonoteca Zoológica, is yet another example of sound collections, being the animal sound library of the Museo Nacional de Ciencias Naturales of Madrid (Spain) [88]. Currently Fonozoo provides about 33,000 metadata records online. The Animal Sound Archive [41] at the Museum für Naturkunde in Berlin presently provides about

120,000 bioacoustical recordings. The Avian Knowledge Network [4] 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 and who (observer)”. For example, at FNJV most of the sounds were recorded by domain experts, who often annotated associated metadata at recording time. On the other hand, some of the other collections cited 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 3.1 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 3.1: 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, Sound file format, Frequency (kHz).

### 3.2.2 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 [59, 45].

Several kinds of projects take advantage of information provided by citizen science and VGI. One example is [71], 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 [68]. 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 format – 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 [16] and the International Union for Conservation of Nature (IUCN) [60].

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.

### 3.2.3 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 [15, 65, 36, 5]. In [15], the authors detect duplicated records in metadata using text distance functions. In [65], the authors use clustering methods and association rules in order to perform data cleaning. In [36], authors improve the quality of relational data using conditional functional dependencies. In [5], 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 is to design algorithms that derive coordinate information from place names [22]. In [99], 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 [34], 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 [55]. Centroid-based approaches, however, may fail to provide the degree of precision 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 as well.

### 3.3 Our Approach

The main idea behind our geographical approach is to contrast geographic distribution maps against the places in which 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

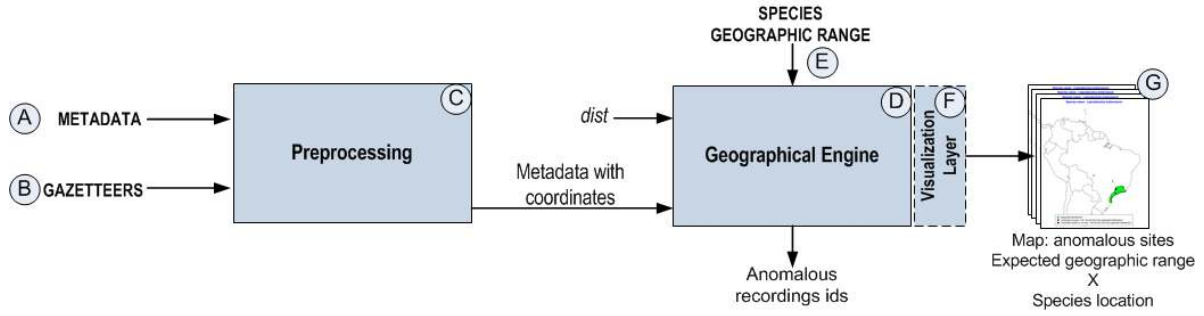


Figure 3.2: Geographical technique to support metadata quality improvement in biological observation databases.

problems are 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 3.2 gives an overview of our approach.

**Step 1 - Preprocessing** The first step of our technique (preprocessing – item C in Figure 3.2) retrieves from metadata both species name  $s$  and the set of places where the species was observed,  $P_s = \{p_1, p_2, \dots, p_n\}$ , 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 function to derive geographic coordinates from gazetteers (item B). Note that, if desired, users can also derive geographic coordinates from some specific georeferencing tool, such as BioGeomancer [48]. 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). Scenario (b) provides polygon coordinates of the region indicated in the metadata. Figure 3.3 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).

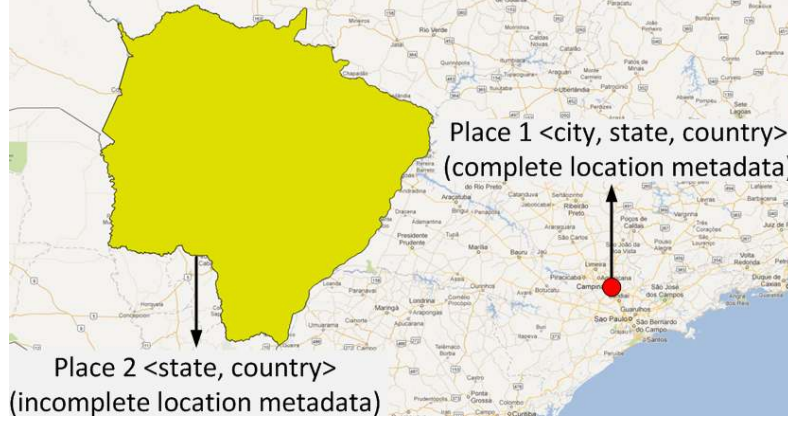


Figure 3.3: Coordinates of two places extracted from gazetteers. Place 1 (a point) is derived from a complete metadata location, containing <city, state, country>names. Place 2 (a polygon) is derived from an incomplete metadata location, containing <state, country>names.

**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. The engine 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,  $GR_s = \{q_1, q_2, \dots, q_m\}$ , 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) [60] or BirdLife International Digital Distribution Maps of Birds [16]. Note that since the geographic range maps play a key aspect in our geographical approach, in order to get better results, the maps must be up to date. Although the geographic range sources cited 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 3.2.2). 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*) 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 3.4 shows the original geographic range  $GR_s$ , the  $dist$  variable set up by the expert, the buffer and the new buffered region  $BGR_s$ .

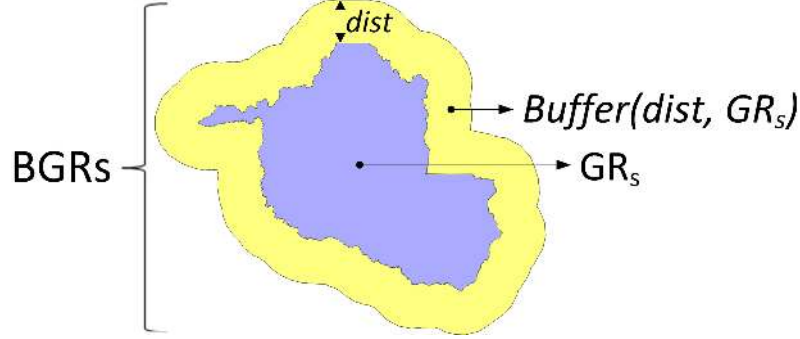


Figure 3.4: A buffered geographic range,  $BGR_s$ , of size  $dist$ .

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. The buffered geographic range is application-dependent and can be arbitrarily large or small. For example, it can be used to cover an area flooded around a lake or river.

Given these definitions, anomalous places are defined to 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:

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

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  $P_s$ , such that  $\lambda_{(BGR_s, P_s)} \subseteq P_s$ .

**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.



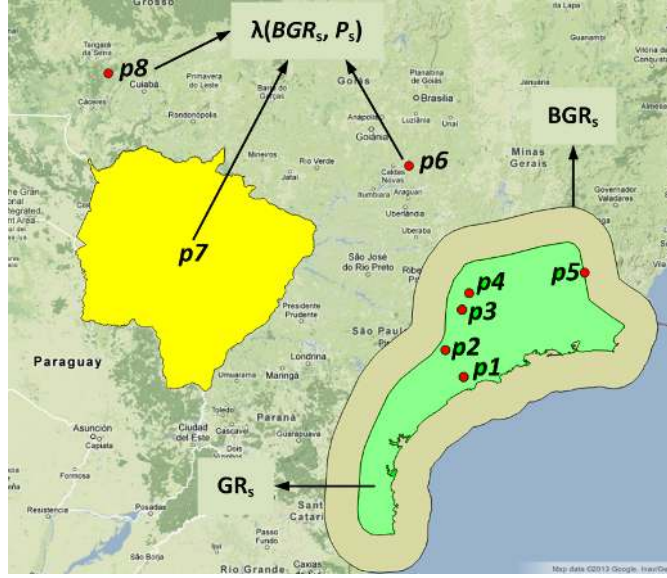


Figure 3.5: Example applied in the technique. The green polygon is  $GR_s$  and the gray polygon is  $BGR_s$ . Places from  $p_1$  to  $p_8$  are the places in which the vocalizations of species *Leptodactylus bokermanni* were recorded,  $P_s$ . The places  $p_6$ ,  $p_7$  (yellow region) and  $p_8$  are anomalous, i.e.  $\lambda(BGR_s, P_s) = \{p_6, p_7, p_8\}$ .

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 3.5 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 is 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.,  $\lambda(BGR_s, P_s) = \{p_6, p_7, p_8\}$ .

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-



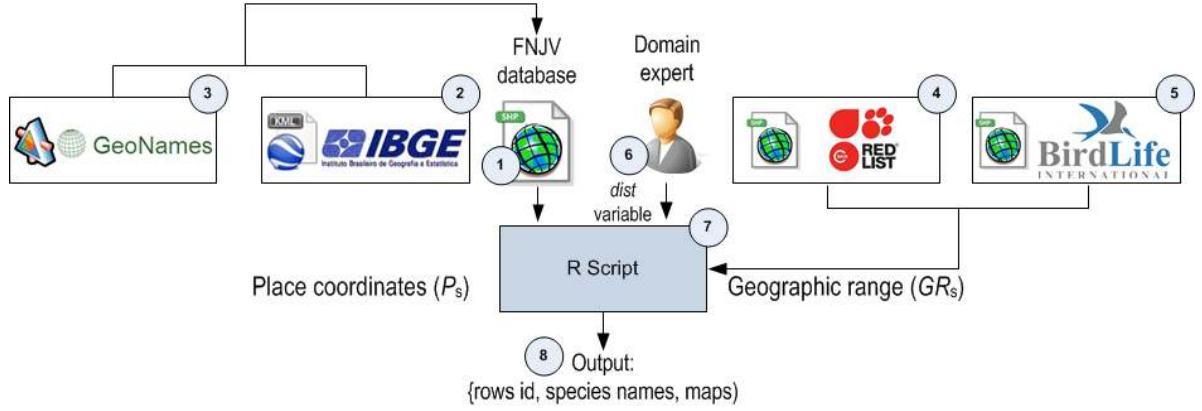


Figure 3.6: Prototype of our Geographical Approach for Metadata Quality Improvement.

automatic, being used to help experts to improve metadata quality. The interaction of experts in part of our process is essential to avoid the detection of false-positive anomalies, since lack of observation records in the geographic range maps does not mean absence of a species.

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.

## 3.4 Case Study

### 3.4.1 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 [96] and the Brazilian Institute of Geography and Statistics (IBGE)[56], using centroids of polygons (cities, states, countries). Note that this methodology may not provide accurate coordinates of the places where the sounds were actually 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 [60] and the BirdLife International Digital Distribution Maps of Birds [16]. These files were adjusted to the WGS84 world geodetic system. Additional map sources can also be used (e.g., National Atlas - Amphibian distribution [8]).

### 3.4.2 Prototype

Our prototype was created using R [57], 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 3.6 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. The 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 exported to the SHP file format.

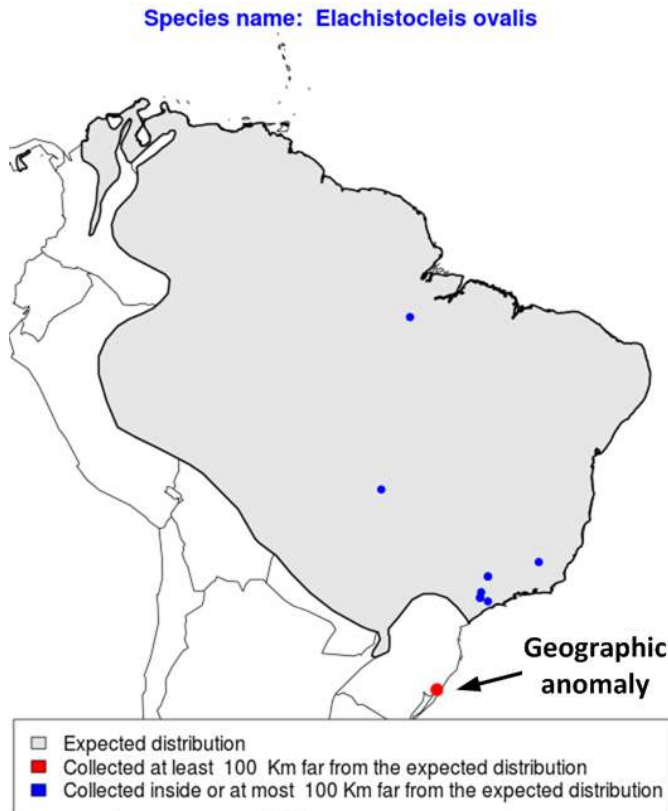


Figure 3.7: Output map generated by the prototype for the *Elachistocleis ovalis* species (Amphibian)– anomaly classified as *outdated metadata*.

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)

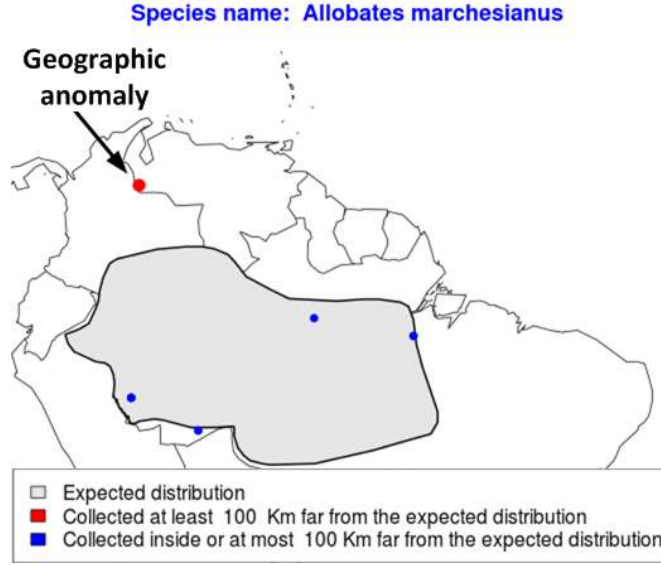


Figure 3.8: Output map generated by the prototype for the *Allobates marchesianus* (Amphibian) – anomaly classified as *error in the distribution range map*.

Visual description: maps containing the rows id, species name, the regions  $GR_s$  where they are expected to live and the places  $P_s$  where the vocalizations were recorded.

Figures 3.7 and 3.8 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  $\in \lambda_{(BGR_s, P_s)}$ .

In contrast, Figure 3.9 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.

### 3.4.3 Results

The prototype was set up with  $dist = 100$  kilometers. Table 3.2 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

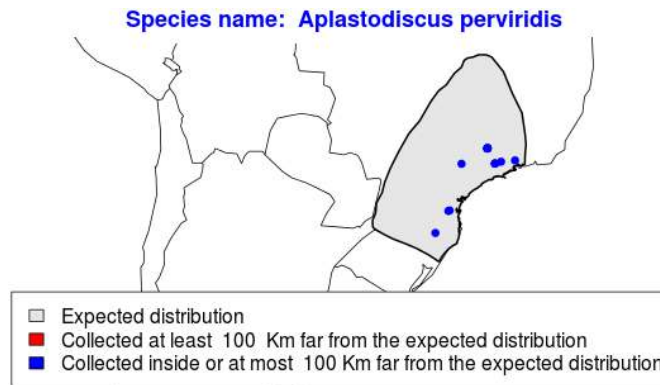


Figure 3.9: Output map generated by the prototype for the *Aplastodiscus perviridis* species, an amphibian species. In this case all observations are non-anomalous.

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 3.2: 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 3.3 details each category.

Figure 3.7 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 3.3, Class B). Figure 3.8 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 3.3, Class C).

Table 3.4 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

Table 3.3: 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.

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 3.3). The fourth column summarizes the corresponding feedback.

Let us clarify the content of Table 3.4 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 anomalies were classified as categories A or D (metadata error or new pattern – according to Table 3.3). 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 3.4 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.

Table 3.4: Biologists feedback for amphibian species observed in anomalous sites.

Species name	Anomalous observations in the database	Class	Comments by Scientists
<i>Allobates marchesianus</i>	1	C	Probably the distribution map is underestimated.
<i>Elachistocleis ovalis</i>	1	B	This name is not valid any longer. Current taxonomic studies indicate that <i>Elachistocleis ovalis</i> should be in fact named <i>Nomen inquirenda</i> [25].
<i>Leptodactylus bokermanni</i>	2	A or D	The points in the middle of Brazil probably are other species. They might be the <i>Adenomera bokermanni</i> .
<i>Pseudis limellum</i>	6	A or D	The point in the Amazon forest region probably corresponds to other species, perhaps <i>Lysapsus limellum</i> .

### 3.5 Discussion

We improve metadata quality through detection of mismatches between location metadata and habitat maps. A mismatch can indicate that either the location metadata or the habitat map is incorrect. It can also indicate new or previously unreported features of species movements, due to, for example, seasonality (e.g., summer, winter), phenology (e.g., due to global warming) or simply a natural variability.

The geographical scale and/or spatial unit needed to study a species' habitat, ecological niche, etc, varies not only with the species, but also with the geographical region and even scope of the research. For instance, some species are restricted to relatively small patches within a region, whereas others, due to roaming habits, food availability, etc, extend across large geographic areas that require large geographical units. Such spatial units also vary with time – and thus a study related to the 1960's may work with larger surfaces than studies based in the 21st century, for the same broader region. Some species may range across continents – and even then, vocalizations may not be the same. Thus, buffer determination must respect not only such issues, but also the kind of study that experts conduct.

Habitat maps are often constructed based on human observations of animal species. However, some species are shy and stealthy, making observations difficult. Thus, lack of observations at a location does not necessarily imply that the location is outside the

habitat. In other words, habitat maps may underestimate the geographic area inhabited by some species.

## **3.6 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.

Since our work requires that scientists manually classify the results provided by our geographical technique, dealing with high volumes of metadata might be highly time-consuming. In order to overcome such limitation, 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 [34].





## Chapter 4

# Adaptive Acquisition of VGI to Fill out Gaps in Biological Observation Metadata

### 4.1 Introduction and Motivation

Biological observation databases contain priceless information that can be used to provide knowledge for broad kinds of research, e.g., global warming, species behavior, food production, etc. Such information can be acquired in many distinct ways and from many distinct sources. In the past, observations were made mainly by domain experts, who needed to spend days, weeks or even months traveling and collecting information in field trips. Such observations were usually made manually, and metadata were annotated in paper, in natural language, most of the time following no predefined data structure. Such metadata may differ according to the domain, but the most important fields are supported by all kinds of observations – i.e., “what” (what was observed), “when”, “where”, “how” (methodology) and “who” (observer). The “where” has become one of the most important pieces of information in observations, since several studies are sensitive to geospatial information.

The way of collecting and persisting biological observation metadata has changed drastically. First, an increasing number of observations started to be provided not only by domain experts involved in projects, but also by non-expert volunteers. Such information, named VGI [59, 45] (Volunteered Geographic Information – See Section 4.2.2), has introduced a new paradigm in the way of collecting data. Second, computational systems with predefined data structures are now commonly used to persist observational metadata. Third, different computational tools have been employed to acquire information, such as web-based forms and mobile applications. In spite of those changes, one feature remains

the same – there is often missing information in the metadata.

Such gaps can bias research into wrong scientific conclusions or even hamper the use of the observations. There are countless reasons for gaps in biological observation metadata – e.g., lack of equipment to measure environmental variables or lack of standards. Gaps can be generalized and classified into two groups: (a) incomplete information in the metadata records and (b) insufficient samples for some specific scenario.

**Incomplete information:** here, not all record fields are filled out, for many reasons, including data capture using heterogeneous devices. For example, when recording animal vocalizations, identification of species is a hard task since many times the scientists do not see the animals whose vocalization they are recording. For this reason, metadata may, for instance, lack a scientific name, providing only partial taxonomic classification. Additional gaps may be caused, for instance, by lack of appropriate instruments at the observation site, or absence of adequate conditions to perform some measurement.

**Insufficient samples:** here, the problem is having an insufficient number of metadata records that fit a scenario. This may occur, for instance, when a team of scientists wants to reuse data collected by another team, but the teams have distinct sampling methodologies. Our work concerns this kind of scenario.

Related work (e.g., [34, 51]) has proposed different strategies that can be used to fill out metadata gaps. Some approaches (e.g., [34]) acquire more metadata records by setting the appropriate information from web sources. Others use experts to manually process the records and fill the gaps. Still others adapt VGI, which is the focus of our work. In [51], for example, scientists create electronic forms to collect metadata from VGI through web-based and mobile applications. However, electronic forms are static and the data acquisition strategy does not consider that metadata requirements may change as researchers acquire more knowledge about some problem. Such dynamicity makes this a challenging problem. Filling out gaps in metadata requires adaptive techniques to collect the right information in the right moment and place. To the best of our knowledge, there is no work that supports such features in our target domain. In order to overcome such challenges, this paper provides a context sensitive approach to collect biological observation metadata using dynamic forms that are updated on the fly by taking into account the metadata being acquired, in real time. Location-sensitive metadata acquisition forms are dynamically adapted according to scientists' requirements.

The rest of this paper is organized as follows. Section 4.2 introduces basic concepts for our approach and related work. Section 4.3 describes our approach. Section 4.4 details our prototype, which collects VGI from mobile devices and sensors over the Arduino platform. Section 4.4 also describes an example using our prototype. Finally, Section 4.5 presents our conclusions and future work.

## 4.2 Background and Related Work

Our research is concerned with using VGI to collect biological observational data. Thus, related work discussed here concerns these issues.

### 4.2.1 Biological Observation Metadata

Biological observations concern the recording of 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)” [23]. Nowadays, observations have been described by several kinds of data types, such as textual information, photos and sounds. Biological collections might concern broad kinds of species, or a specific category thereof. For example, the Cornell Lab of Ornithology [29] contains sound recordings, photos and associated metadata restricted to birds [21]. The Fonoteca Neotropical Jacques Viellieard (FNJV) [39] has sound recordings and associated metadata of all vertebrate groups (fishes, amphibians, reptiles, birds and mammals) and some groups of invertebrates (as insects and arachnids). The plant database of the Kwantlen Polytechnic University [67] provides images of several species of plants and their associated metadata.

Biological observation metadata can be provided in two ways. Historically, observers went to the field, performed their observations and annotated the observational information in paper; eventually, observations were transcribed to spreadsheets. Here, the number of observations was low, since observations were mostly provided by a limited number of domain experts. Today, observational information is often acquired using mobile devices and sensors. The number of observations performed is much higher, and there is an increasing number of volunteers who contribute with their observations.

Metadata must be curated, because there are several quality issues, including incomplete and incorrect information. For instance, acquiring observational information from volunteers may affect the quality of the metadata provided, since volunteers are not always domain experts. There are approaches that overcome such problems, by focusing on quality aspects of the metadata, e.g., [32, 73, 72]. Another problem is related to the fact that there is often missing information in such metadata. Distinct approaches have also been proposed for this problem, each of which focusing on different aspects, e.g., [34]. Moreover, biological databases are always changing. As new observations are provided, gaps in the metadata are filled or new gaps arise. Filling out such gaps is an open research problem.

### 4.2.2 Volunteered Geographic Information (VGI)

Citizen science is the term often used to describe communities or networks of citizens who act as observers in some domain of science. Analogously, Volunteered Geographic Information (VGI) refers to data provided by citizens, in particular, including geographic information [59, 45]. The use of VGI has been shown to be successfully used by an increasing number of projects. There are different projects and initiatives, in different domains of science, boosting the number of observations in their collections by using VGI.

In the biology domain, for example, there is the Christmas Bird Count project [68], that is related to animal preservation and environmental studies. This project is an effort to perform a mid-winter census of bird populations in the North America. This kind of project considers, among others, information provided by citizen science and VGI to create geographic distribution range maps for several species. Another project [14] uses volunteers to record and map the 13-year periodical cicada in South Carolina/USA. Observations are entered through websites of the University of South Carolina<sup>1</sup> and the South Carolina Forestry Commission. Still in the domain of cicada observations, The Cicada project [27] provided schematics to assemble sensors to measure soil temperature. This temperature is a key aspect on analyzing periodical cicadas. The temperature information is submitted through websites, in predefined forms. Another project is the Canada Nature Watch Programs [95], which receives observations from volunteers for several domains (e.g., plant phenology, worms, frogs, and ice). One possible way to send an observation is by printing a form, filling it out manually and sending it through mail to the project committee. Their purpose is to study the effects of climate change and other impacts on biodiversity.

There are also examples of VGI in non-biological domains. For instance, the Open Street Maps project [50] provides editable maps of the world by using information from volunteers. In the Geo-Referenced Field Photo Library project [98] researchers created a mobile app in which volunteers use smartphones to take geo-referenced photos in the field in an effort to document observations of landscapes, agriculture, forests, natural disasters, and wildlife. Another example [71] uses citizens to measure people's 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. In [70], location based social networks are used to acquire spatio-temporal data on forest fires.

The similarity among such projects is the fact that all of them are VGI powered. Their differences come from the variables recorded and the way they collect the data,

---

<sup>1</sup><http://cricket.biol.sc.edu/cicada/contact.html>

which varies from simple printed forms sent by mail to tailored mobile software. They indicate the necessity of systems to support acquisition of VGI data in a wide range of domains.

### 4.2.3 Approaches to collect VGI

In the past, before the advent of personal computers and smartphones, VGI was collected mainly through the use of printed forms, created by researchers. After volunteers filled out the forms, they were usually returned to the researchers via mail. The whole process of creating, deploying and receiving the forms used to take several days or even months. Besides, depending on the number of forms received, it was hard to organize and process the information. Even nowadays, it is still possible to find projects that collect VGI using the same strategy, differentiated only by the fact that forms might be returned, in most cases, by both mail or email. An example is the MeadoWatch project at the Mount Rainier National Park (USA) in which the goal is to study climate change through observations of plants [75].

According to [52], two basic technologies are responsible for boosting the acquisition of VGI: geo-referencing and the Web 2.0. GPS devices and GPS-equipped smartphones easily enable geo-referencing and the Web 2.0 enables collecting information generated by volunteers, automatically persisting all information in databases with semantics. These technologies are the source of two popular ways of collecting VGI: (a) forms published in websites and (b) forms published in mobile applications.

In the first strategy (a), programming skills are required to create the websites with the electronic forms. Besides, creating such websites is time consuming. Some tools enable easy creation of electronic forms; however, such tools are limited when providing ways of personalizing the forms. The Nature's Notebook project [81] uses the strategy of web forms. Here, volunteers provide observations of plants and animals and all metadata is submitted in an electronic form embedded in the website. A hurdle in such strategy is the fact that such forms are static, requiring spending time changing the web site if there is a need to ask volunteers other questions.

The second strategy (b) differs basically by the flexibility that mobile devices provide. Here, volunteers can go to the field and fill out the forms directly in the mobile devices. Another advantage here is that volunteers can use embedded sensors (e.g., GPS and camera) in the mobile devices in order to feed more data. However, in this strategy the forms are also static, requiring time to program and update the mobile applications to change the forms. One example of a project that uses such a strategy is the Creek Watch project [63]. It allows volunteers to report information about waterways, to support water management programs.

There are also approaches that support the creation of personalized questionnaires to collect VGI, such as the Open Data Kit [51], wq [86], Sensr [62] and CitSci.org [80]. However, none of them enables performing dynamic changes in the questionnaires by considering the gaps that may occur in the metadata as new VGI information is acquired or as researchers evolve their knowledge about some problem. Moreover, there are no dynamic changes in the strategy of where to send volunteers. Our approach differs from other approaches by considering the necessity of such dynamism, as detailed in the next section.

## 4.3 Our Approach

The goal of our approach is to provide a strategy in which researchers take advantage of VGI to fill out gaps in biological observation metadata. As discussed in Section 4.1, there are two kinds of gaps in biological observation metadata: empty fields in one or more records, or absence of records. We are concerned with this second kind of gap.

Our strategy considers that the metadata fields are fixed and predefined. This assumption ensures that old records will not be invalidated by changes in metadata. We also consider that VGI will be used to provide additional records for situations in which the database already has records from previous projects. In other words, VGI is used to add new records to curated collections of observational metadata (as opposed to their use to create such a collection from scratch). Our approach is context sensitive, i.e., researchers can define on the fly the VGI data that must be acquired. In order to improve the effectiveness of our approach we also support the definition of spatial regions where an information must be acquired.

### 4.3.1 Overview

Our approach is divided in three steps, as shown in Figure 4.1.

**Step 1 - Detecting gaps in the metadata.** Here, scientists will use a suite of algorithms for gap detection (item B in Figure 4.1) in a biological database. These algorithms detect gaps in the metadata based on both scientist and research requirements. To do so, algorithms use as input parameters provided by the scientists, as well as the biological data and associated metadata records. The suite of algorithms detects gaps by considering the most important attributes in biological observation metadata, i.e., **what** (the species observed), **where** (location of the observation), **when** (date/time of the observation) and **how** (the methodology used). The algorithms provide the gaps to step 2.

**Step 2 - Acquiring VGI.** Here, (item C, Figure 4.1) scientists analyze the gaps

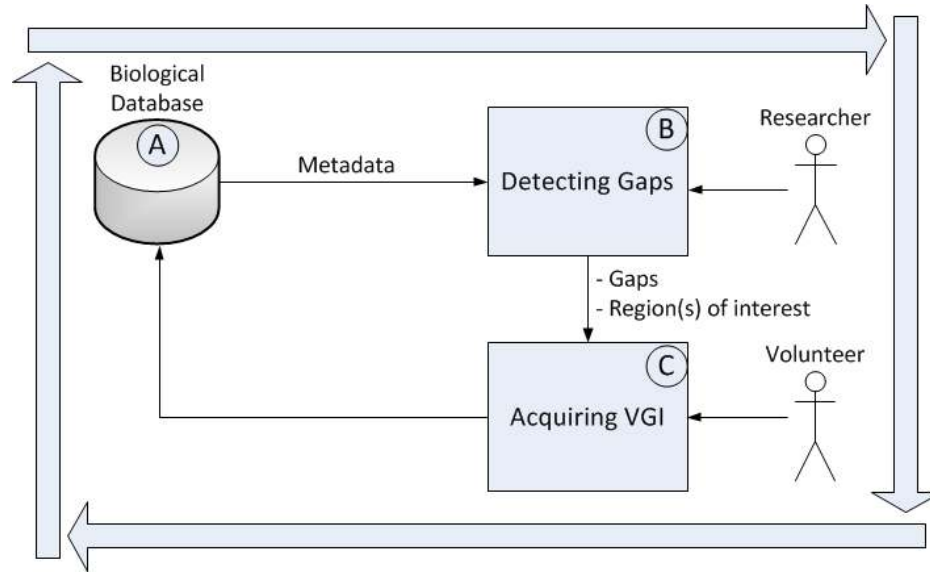


Figure 4.1: Our approach for adaptive acquisition of VGI to fill out gaps in biological observation metadata.

provided by step 1, to determine what kinds of VGI should be acquired. As a result, researchers must define (a) a spatial area of interest where VGI must be acquired (*where*); (b) the focus of the VGI (i.e., the target species, *what*) and (c) extra requirements (e.g., observations must be performed in a given time of the day - *when* and/or *how*). Since the database fields may include several kinds of data types, VGI acquisition is not restricted to only textual information, comprising other data types, such as video, audio, etc. VGI metadata fields can be entered by humans and at the same time complemented by sensors.

**Step 3 - Iterating the process.** Metadata requirements might change as researchers acquire new knowledge about their problems. Biological databases are always changing, with new records inserted in the collection, e.g., when experts detect the need to collect new data. This scenario guided us to include a cycle in our approach, creating an interactive process in which researchers can change *on the fly* any of the VGI collection directives (i.e., (a), (b) or (c) in step 2). This is the most important part of our contribution, since other VGI approaches do not support on-the-fly gap specification.

### 4.3.2 Detection of Gaps

#### Gaps in spatial coverage.

In order to illustrate the applicability of our approach, consider an example in which a researcher is studying the effect of a disease  $D$  that is infecting the species *Elachistocleis ovalis* (an amphibian species). To do so, (s)he needs to have at least  $n$  observations of

such a species with spatial coverage  $S$  to perform experiments. In order to check if this condition is satisfied (i.e., check if there is a gap – **Step 1**), the researcher firstly narrows the search space by selecting database records of interest. Afterwards, (s)he executes algorithms to detect spatial regions in which there is a low number of observations for the target species. If this spatial gap is detected, the researcher defines, for example, the following scenario to acquire VGI (**Step 2**):

- Target = Species *Elachistocleis ovalis*
- Region of interest = a circular area defined by the point(-30.118255,-50.422468) with a 10 kilometers radius
- Metadata field 1 = Time and date of the observation
- Metadata field 2 = Air temperature
- Metadata field 3 = Air humidity
- Metadata field 4 = Picture of the frog
- Metadata field 5 = Location of the observation (coordinates)
- Metadata field 6 = Number of individuals of the same species seen close to the individual observed

The scientist uses this information to configure a web form that is sent to volunteers to collect these data, for that spatial coverage.

After receiving VGI containing the pre-defined metadata fields, all observations are integrated with the biological database (item A in Figure 4.1), increasing the samples for such species in the target region. Note that our approach enables location-sensitive acquisition of information, by defining the region of interest where VGI must be collected. After integrating the new observations with the previous records, researchers can perform new data analyses. In order to improve the analyses, curation of VGI must be performed by the domain expert, since VGI data quality may vary broadly. Strategies to deal with data quality can be used, such as [89, 73, 72].

Figure 4.2 shows the spatial coverage  $S$  that, in our example, refers to the circle defined by point(-30.118255,-50.422468) with radius of 10 kilometers. All VGI in the database ( $O_n$ ) were expected to be performed within the region  $S$ . Blue points (from  $O_1$  to  $O_{10}$ ) are locations where observations were performed within region  $S$ . Opposite to the blue points, red points (from  $O_{11}$  to  $O_{17}$ ) were observations performed outside the expected region  $S$  and the expert must decide if such observations are going to be inserted in the database or not.



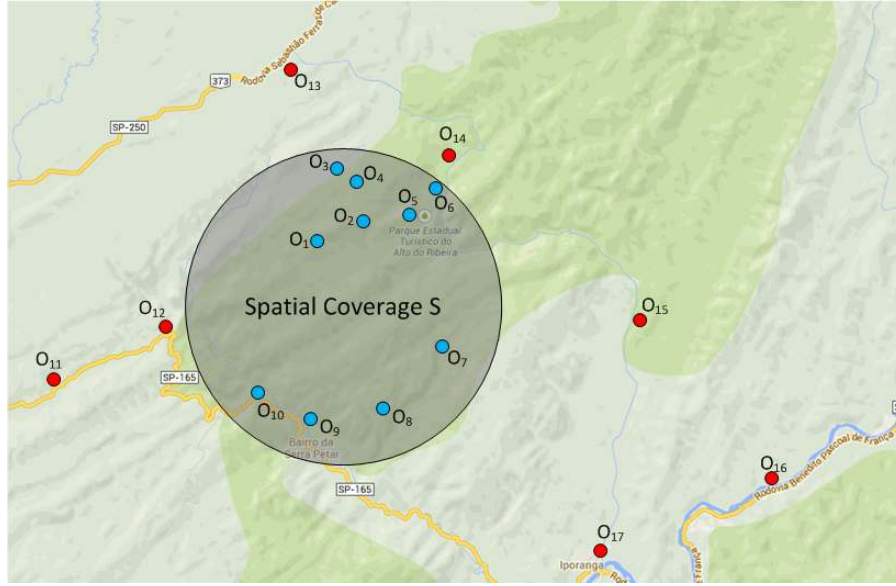


Figure 4.2: Observations  $O_n$  provided by VGI. Blue points are observations provided within the expected spatial coverage  $S$ . Red points are out of the spatial coverage.

Note that, in this example, just one set of database fields compose the form that guided the acquisition of VGI in the spatial coverage  $S$ . However, it could be also possible to define different forms to be deployed in different regions of interest simultaneously.

In this example, the gap detection algorithm detected a gap by checking a **where** field (the spatial coverage). However, the suite of algorithms for gap detection can also detect gaps by considering **how**, **what** and **when** fields. Here follows an example of *how* (gaps in methodology), *what* (gaps in species observed) and *when* (gaps in temporal coverage) scenarios that typical algorithms must consider.

#### Gaps in methodology.

Methodology gaps can be divided into two scenarios. **Scenario 1:** A researcher needs observations performed under a specific environmental condition. For example, all observations of species  $s$  were made in rainy days and the researcher needs samples in sunny days. **Scenario 2:** A researcher needs animal sound recordings made with the equipment  $E$ , brand  $B$ . Depending on the equipment used, the quality of animal sound recordings might be affected.

Let us give an example for *Scenario 1*. Consider that the researcher acquired enough observations for species *Elachistocleis ovalis* using VGI. Also consider that after the acquisition of such observations the researcher detected another gap (using the suite of algorithms), by checking the absence of observations of species *Elachistocleis ovalis* in days with air humidity higher than 65%. The researcher could start over by defining the

same database fields to compose the questionnaire to collect VGI. However, here (s)he would change a question's priority order by defining the first form questions as those that represent the most important gaps in the database (**Step 3**), e.g., defining the *air humidity* as the first form field, guiding volunteers towards focusing firstly on the most important information.

#### **Gaps in species observed.**

Consider that the researcher needs observations of species *Aplastodiscus perviridis* and *Allobates marchesianus* (another kind of amphibians). Here, a *what* gap might be absence or low number of observations of such species. The *what* gap can also take advantage of a *where* analysis, by contrasting *species* against *number of observations* and against *location of observations*, e.g., showing which species are missing in a sample within a region.

#### **Gaps in temporal coverage.**

Consider that the researcher needs observations of amphibian species *Aplastodiscus perviridis* and *Allobates marchesianus* within a specific time frame. For example, there are no observations of such species between 1AM and 3AM. Here, similarly to the “*what*” example, it is also possible to take advantage of the *where* analysis by analysing *when* and *where* in conjunction.

For all such scenarios, our context sensitive approach enables researchers to change on the fly the list of data they want to acquire, by providing a never ending cycle that is adapted to their current requirements. Such requirements will be in turn affected by the metadata collection evolution.

## **4.4 Prototype**

We developed a prototype to enable adaptive acquisition of VGI, i.e., for item C of Figure 4.1. Our VGI comes from two kinds of sensor sources: human “sensors” and other (standard) sensors. Human sensors are citizens that enter data on Android-platform devices, while other sensor-based information is collected using the Arduino platform.

### **4.4.1 The Arduino Platform**

In the past, using sensors to collect data was not a simple task, since it required the development of specific hardware by people with high skills in electronics. This could represent a bottleneck in the development of projects, by consuming a considerable amount of the project's time and money. In the course of time, cheap programmable electronic boards have been created, accelerating and facilitating users to interface with different kinds of sensors.

There are now several kinds of such boards, each of which with distinct features. For example, Raspberry [66] and Beaglebone [12] are credit-card-sized single-board computers, with proprietary hardware, that usually run Linux-based operational systems and that provide GPIO (General Purpose Input/Output) pins to interface with a broad kind of devices, including sensors.

The Arduino board [10] is a single-board microcontroller for electronic prototyping, made with an open-source hardware. It also provides I/O pins that enable communication with several kinds of devices (e.g. LEDs, servos, ethernet boards, etc.) and sensors (air pressure, air temperature, luminosity, motion, RFID, etc.). Furthermore, Arduino is cheaper than single-board computers, and its easier to use when interfacing with devices and sensors. Arduino can be programmable in C or C++ languages.

There are plenty of add-on hardware modules for Arduino, known as *shields*, which are expansion boards that plug into the Arduino pin-headers. Such shields are one of the features that make Arduino so attractive and broadly used. Shields can provide GPS, ethernet, LCD display, SD card interface, standardized connection pins, etc., letting users free of spending time on creating their own hardware. Another useful feature is that several kinds of sensors are provided using the Grove system. Grove is a ready-to-use tool set, similar to LEGO. It takes a building block approach to assemble electronics, simplifying their use. The Grove system is composed by a base shield and several modules with standardized connectors. Our project is focused on providing an environment in which scientists do not need to have skills in programming and/or electronics. Thus, we decided to use Arduino in conjunction with Grove shields and Grove sensors, in order to simplify and enable the use of sensors to collect VGI. Our prototype only requires that scientists assemble the hardware (similar to constructing a LEGO toy) and set up a XML file, as described on Subsection 4.4.3.

#### 4.4.2 Defining on the Fly Which Data Fields Must be Acquired

Our prototype is supported by a local web server that does not depend on third-party services. Therefore, we ensure that the service will never be stopped by changes that might occur when using remote third-party services. Figure 4.3 shows an overview of our prototype, which enables collecting data from both humans and sensors. Collecting data from sensors and humans has different features, as follows.

***Humans as sensors:*** humans can act as sensors to provide VGI by observing things or phenomena. Our prototype supports the whole process of observation, that includes (a) remotely configuring the form fields that will be presented to the volunteers, (b) publishing the form in mobile devices and (c) collecting and persisting the VGI. In (a), the form definitions must be set up through a XML file that is defined by experts. We

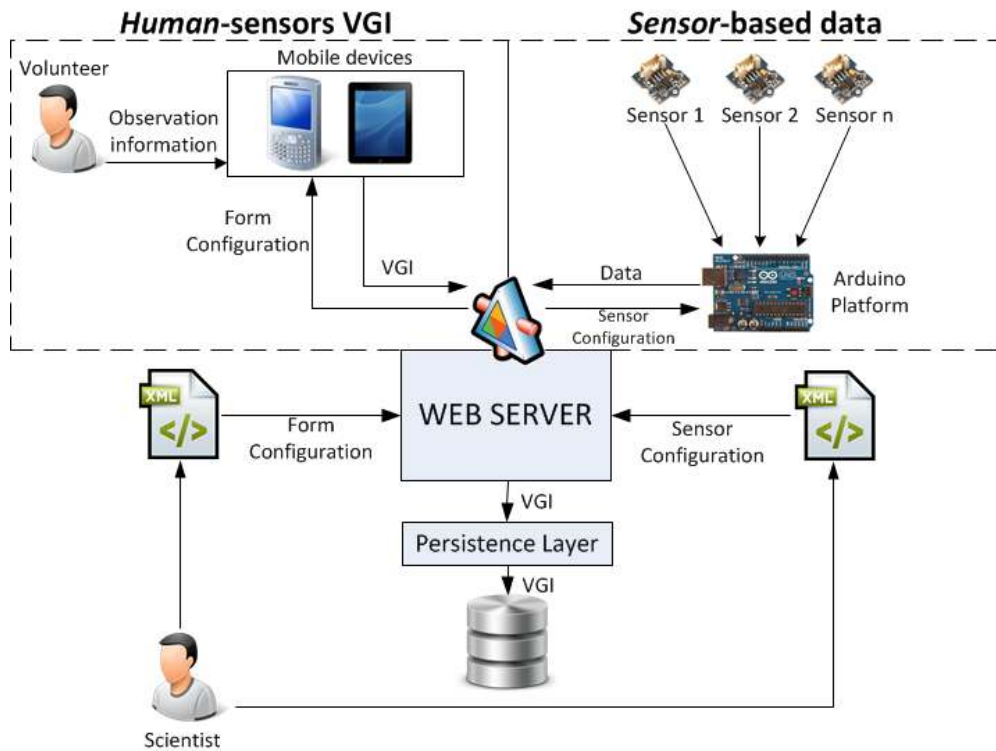


Figure 4.3: Scientists remotely configure forms and sensors, defining which attributes to collect.

chose to use XML because it is a consensual standard format, and easier to integrate with other sources.

In our XML files, experts define the order and the fields that must compose the electronic form, e.g., location where the observation is performed, date and time of the observation, picture of what was observed and name of the species observed. Here XML files can be shared, enabling reuse of such configurations as templates. The form is made available by our web server through web services (developed in Java + Apache Tomcat server). Any mobile device using our mobile application (developed in the Android platform) can connect into the web service, acquire data to configure the form, and create an electronic form dynamically. Afterwards, volunteers can record the observation, by filling out the form using mobile devices. Geographic coordinates and photos are taken directly from the GPS and camera embedded into the mobile device. All VGI is returned through the web service. Note that pictures are serialized before being submitted. Finally, the web server receives the VGI and persists all information into a CSV file. We chose persisting VGI in CSV files because any software to manage spreadsheets can open CSV files. CSV files can also be easily imported into any database. Since spreadsheets are a major tool for scientists who deal with observations, this is a good choice.

**Sensors as data sources for VGI:** similarly, our prototype also enables collecting data from sensors using the Arduino platform. To do so, we developed a framework that enables easy configuration of Arduino. Our framework allows experts to (remotely) configure a set of sensors, such as GPS, air temperature, air humidity, gas and light sensors. Configuring the Arduino platform through our framework requires experts to create a XML file, defining the sensors that will be used. The whole process (from defining the information to acquire, to collecting and persisting such data) follows the same procedure as the topic *humans as sensors* for VGI.

Since our prototype is configured through a XML file, whenever a researcher detects a different gap, the XML file can be updated to reflect the current research requirements.

### 4.4.3 Example

In order to illustrate the use of our prototype, recall the scenario of Section 4.3, in which a researcher is studying the effect of a disease  $D$  that is infecting the frog species *Elachistocleis ovalis*. The expert is interested in performing a spatial analysis to detect “where” gaps. Here, step 1 is performed by running a gap detection algorithm [32]. This algorithm is based on a geospatial approach that contrasts geographic distribution maps (regions in which species are expected to live) against the places in which the observations were made (as per location metadata). Figure 4.4 shows a gap detected by our algorithm for records of the species *Elachistocleis ovalis*. Our input are the 30,000 species observation records of the FNJV repository [39].

The light gray polygons represent the regions ( $R$ ) where the species is expected to live. Blue points are locations where observations were performed inside or at most 100Km far from the expected regions ( $R$ ). The red point represents a location where the observation was performed farther than 100Km from the expected regions ( $R$ ). The buffer (100Km) is adjustable and is set up by the domain expert. Note that in Figure 4.4 the red point represents an outlier (anomaly). In this example, such an outlier is considered a *spatial coverage gap*, since the red point evidences the presence of such a species in a region where there is just one observation. Therefore, here the gap is the low number of observations of the species *Elachistocleis ovalis* in a region, i.e., **what** (the species) and **where** (the location where there is a low number of observations).

After detecting the gap, the expert must first define the region where VGI must be acquired. Such a region can be specified in our prototype by defining a point coordinate and a radius. Afterwards, the researcher must define the data fields that will be presented to the volunteers as well as their respective data types. Listing 4.1 shows a XML configuration file where this information is defined. This XML file will configure forms that will be shown in smartphones and tablets. Note that Listing 4.1 defines the

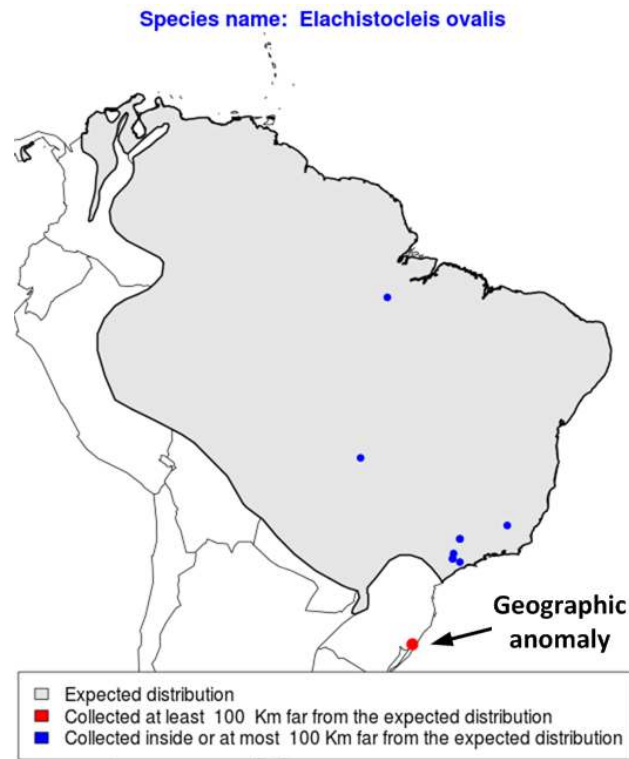


Figure 4.4: Output map generated by the gap detection algorithm for the *Elachistocleis ovalis* species (Amphibian). The anomaly (red point) led to the spatial gap detection.

questions that must be deployed, in conjunction with their respective label, field name that will be associated with the data and the data type. It also defines the region where the questionnaire must be deployed. For instance, line 3 defines the target species. Lines 13 to 18 request information on close-by individuals. Lines 4 to 8 define the region of interest, and so on.

Listing 4.1: XML configuration file to create the form to collect VGI

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <vgi source="mobile">
3   <field type="target" field_name="species_name" label="
      Elachistocleis ovalis."/>
4   <field type="spatial_coverage_definition">
5     <lat>-30.118255</lat>
6     <lon>-50.422468</lon>
7     <radius>20</radius>
8   </field>
9   <field type="string" field_name="air_temperature" label="Inform
      the air temperature."/>

```

```

10  <field type="string" field_name="observer" label="Inform the
    air humidity."/>
11  <field type="picture" field_name="species_picture" label="Take
    a picture of the species"/>
12  <field type="location" field_name="location" label="Get the
    location of the observation"/>
13  <field type="radiobutton" field_name="number_of_individuals"
    label="How many individuals of the same species were seen
    close to the species observed?">
14    <item>1</item>
15    <item>2</item>
16    <item>3</item>
17    <item>More than 3</item>
18  </field>
19  <field type="time" field_name="observation_local_time"/>
20  <field type="date" field_name="observation_local_date"/>
21 </vgi>

```

---

The same kind of configuration must be used when acquiring data from sensors, except that additional information about sensor configuration must be informed. In our prototype, to collect data from sensors, we have used one Arduino board, one GPS sensor, one Ethernet shield, one Grove base shield and one humidity/temperature sensor. Figure 4.5 shows the hardware we have used to collect air temperature and humidity using the Arduino platform.

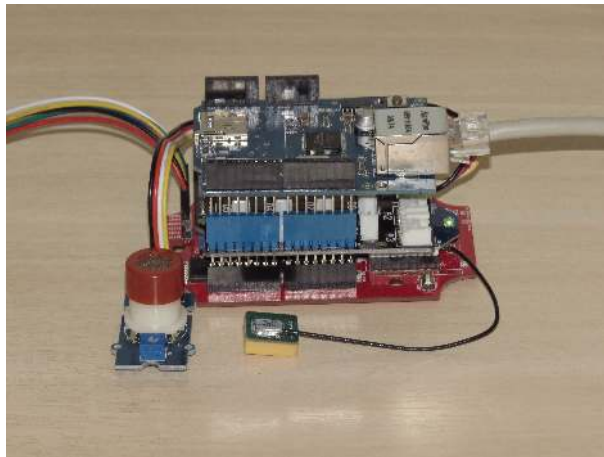
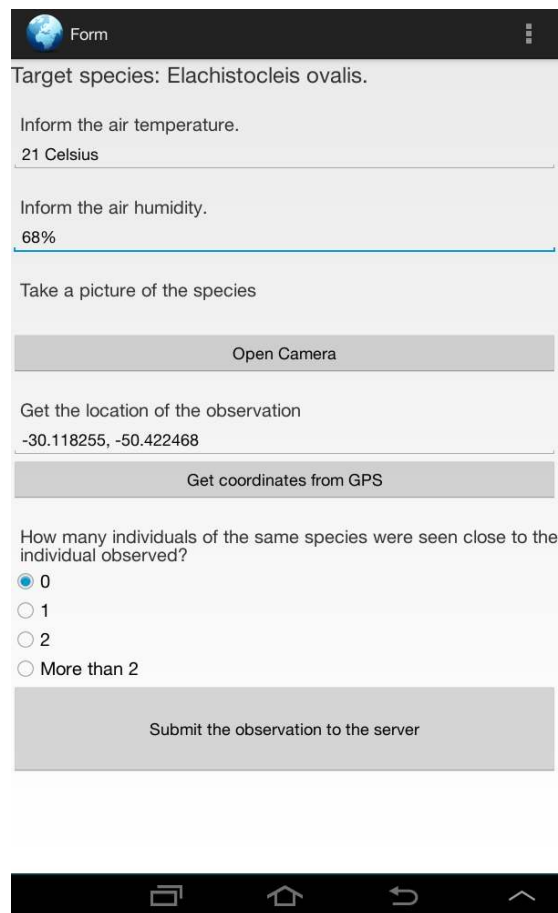


Figure 4.5: Hardware used to collect *Sensor-based VGI*.

The XML configuration file must be saved in the web server where the server application runs. Each volunteer who wants to provide observations by using our mobile

platform must install our mobile application prototype, open it and introduce the IP of the web server. After that, the dynamic form is created in the client side, based on the XML configuration file. Figure 4.6 shows a screenshot of the form created in the mobile device, based on the XML configuration file defined by the researcher. In the current version of our prototype, all observations are saved in a CSV file for future integration with the original database. Future versions of our prototype will integrate VGI data into the database automatically. The main barrier to automatic integration is data quality and curation.

Our prototype is available on the internet<sup>2</sup>. Its current version enables defining one region at a time to collect VGI. Future versions will comprise several regions at the same time, including the definition of distinct lists of metadata attributes to be acquired in each region.



The screenshot shows a mobile application interface titled "Form". It contains several input fields and buttons. The first field is "Target species: Elachistocleis ovalis." followed by "Inform the air temperature." with the value "21 Celsius". Below that is "Inform the air humidity." with the value "68%". Then there is a section "Take a picture of the species" with an "Open Camera" button. This is followed by "Get the location of the observation" with the coordinates "-30.118255, -50.422468" and a "Get coordinates from GPS" button. The next section asks "How many individuals of the same species were seen close to the individual observed?" with radio button options for "0", "1", "2", and "More than 2". The "0" option is selected. At the bottom is a large "Submit the observation to the server" button. The Android navigation bar is visible at the very bottom.

Figure 4.6: Screenshot of our Android application prototype. The form is created on the fly, based on the XML configuration file.

---

<sup>2</sup><http://www.ic.unicamp.br/~dcintra>



## 4.5 Conclusions and Future Work

We presented a context sensitive approach to collect VGI in order to fill out gaps in biological observation metadata. Our approach considers the dynamics of changing needs in research requirements, by assuming that *gaps* may change as new observations are made or as researchers have new insights about some problem. Our interactive location-sensitive approach enables researchers to define on-the-fly which information is required and the regions where information must be collected.

As future work we envisage to focus on (a) improving our prototype by enabling the deployment of forms in regions with arbitrary shapes instead of regions defined by circular areas; (b) enabling the customization of deployment of forms in different regions at the same time; (c) automatically persisting the information acquired into the original database and (d) facilitating the creation of the XML configuration files, by creating a visual tool for selection of the database fields. Another open issue is how to figure out the most accurate information upon conflicting volunteers data.



# Chapter 5

## Conclusions and Extensions

In this thesis we focused on overcoming some problems related to the management of biological observation databases, tackling mainly the problems of quality in metadata. Our motivation came from challenges faced by biologists on managing large amounts of animal sound recordings, using the Fonoteca Neotropical Jacques Vielliard as a real world case study.

We provided strategies to detect and fix some categories of *anomalies* (metadata error, outdated metadata, errors in species distribution range maps, possible new species pattern) and some kinds of *gaps* in metadata (missing attributes and lack of records fitting a specific scenario).

The web site we developed to share animal sound metadata has been accessed by users from 75 countries. Besides, this work was mentioned in a Brazilian TV news program [92], in the UNICAMP newspaper [7] and in an electronic article published by FAPESP [6].

### 5.1 Main Contributions

The contributions of this thesis can be grouped into two categories: (i) *Detecting and fixing anomalies in metadata* and (ii) *Detecting and fixing metadata gaps*.

**Detecting and fixing anomalies in metadata.** Chapter 3 provides an approach to detect anomalies in metadata. Here, our contribution is the geographical analysis algorithm, which detects metadata anomalies that are only possible to be detected when considering geographical attributes.

**Detecting and fixing metadata gaps.** Both Chapters 2 and 4 provide contributions in this category. In Chapter 2, our contribution is the architecture for detecting and enriching missing environmental variable attributes in the metadata. The main point

in the architecture is the query mechanism, that has two roles. First, it fills metadata gaps (missing attributes). Second, it provides information that goes beyond the metadata attributes, providing additional contextual information to observations. In Chapter 4 the contribution comes from the context sensitive approach to collect VGI in order to fill out gaps in metadata. Here, our main contribution derives from the fact that the notion of *gaps* may change as new observations are made or as researchers have new insights about some problem. Our approach takes this dynamicity into consideration.

In both cases, we provide results not yet tackled by related work. For example, related work for *detecting and fixing anomalies in metadata* does not adequately consider a variety of issues concerning such metadata, e.g., misnaming of species, location uncertainty and imprecision concerning where observations were recorded. Even when they tackle some of such issues, they do not perform spatial analysis, ignoring important and hidden aspects of biological observation metadata. Related work for *detecting and fixing metadata gaps* does not consider the dynamicity of changes of metadata requirements.

Summarizing, our contributions are:

- Definition of a geographical approach for detection of anomalies in biological observation metadata;
- Specification of an architecture for retrieval of animal sound recordings. This architecture provides a query mechanism that fills out gaps in environmental context variables by using external data sources;
- Validation of the above by actual implementation of a prototype that runs over our web system to share biological observation metadata;
- Specification of a methodology to fill out gaps in biological observation databases by acquiring information from sensors and VGI provided by dynamic questionnaires;
- A framework that enables adaptive acquisition of VGI using mobile devices and sensors;

## 5.2 Extensions

There are many possible extensions to this work. Examples of some of these extensions are:

- Investigate additional rules/algorithms to derive context variables, in order to come up with a full-fledged context ontology to enlarge the possibilities of derivation of context variables (see Chapter 2);

- Design and develop parallel data collection procedures, to allow cross validation of filling missing data algorithms. This would also be used to better validate the framework of Chapter 2;
- Investigate the use of distribution polygons instead of polygon centroids in the algorithms of Chapter 3;
- Design mechanisms to find out appropriate external context data sources to support the derivation of context variables (see Chapter 2);
- Design algorithms to calculate imprecision in the process of derivation of context variables (see Chapter 2). Since regions in which recordings took place do not necessarily apply to a specific spot, the values derived may be imprecise. This imprecision might be informed to researchers, and query processing might take this imprecision into consideration;
- Design mechanisms to share copyrighted animal sounds (or other kinds of observation content) with biologists who want to use them for research. To do this, it is necessary to provide distinct access control mechanisms in our web site (Chapter 2). This control may be also exercised for all downloads, and will create new data files that can be further used for data mining. For example, scientists can find other scientists that are working with the same species calls, improving interaction in the scientific community;
- Employ supervised learning algorithms to reduce the burden of scientists on classifying metadata anomalies in the geographical approach for metadata quality improvement (see Chapter 3). The supervised learning algorithms might be used to recommend anomaly classes to be reviewed by scientists;
- Integrate the query mechanism from Chapter 2 with the geographical approach from Chapter 3 in order to consider enriched environmental variables in the spatial analysis;
- Develop a graphical user interface (GUI) to facilitate the creation of the XML configuration files in our adaptive framework to fill out gaps in biological databases (see Chapter 4);
- Improve the framework of Chapter 4 to fill out gaps by enabling the deployment of forms in regions with arbitrary shapes instead of regions defined by circular areas. Moreover, allow interface customization - e.g., so that scientists can enter arbitrarily shaped regions;

- Integrate implementations into the framework we have developed as a web site system, in order to provide a full-fledged system to manage recordings and metadata from biological observation databases. For instance, anomalies detected (Chapter 3) can be used to direct VGI work (Chapter 4);
- Consider the curation of a backlog of demand for VGI, which can be met according to volunteers' location and availability. An example of such backlog could be to improve georeferencing of old metadata records.

# Bibliography

- [1] Encyclopedia Britannica - Academic Edition. Available: <http://www.britannica.com/EBchecked/topic/409141/Neotropical-region> (Accessed on 11/2011).
- [2] Incorporated Research Institutions of Seismology (IRIS). <http://www.iris.edu/>.
- [3] Laboratory of Information Systems - Institute of Computing, UNICAMP. <http://www.lis.ic.unicamp.br>.
- [4] AKN. Avian Knowledge Network. <http://www.avianknowledge.net> (Accessed on 12/2012).
- [5] A. Alabri and J. Hunter. Enhancing the quality and trust of citizen science data. In *IEEE VI International Conference on e-Science*, pages 81–88. IEEE, 2010.
- [6] E. Alisson. Projeto recupera acervo da Unicamp com sons dos animais. Available: <http://agencia.fapesp.br/16760>. Agência FAPESP, Jan 2013.
- [7] S. Anunciação. Cantos e outros sons para todos os cantos. *Jornal da UNICAMP*, Year XXVI(528):5, May 2012.
- [8] National Atlas. Amphibians distribution. <http://www.nationalatlas.gov/mld/amphib.html> (Accessed on: January, 2013).
- [9] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.
- [10] M. Banzi. *Getting Started with Arduino*. ” O’Reilly Media, Inc.”, 2009.
- [11] R. Bardeli. Similarity search in animal sound databases. *IEEE Transactions on Multimedia*, 11(1):68–76, 2009.
- [12] S. Barrett and J. Kridner. *Bad to the Bone: Crafting Electronic Systems with BeagleBone and BeagleBone Black*, volume 41. Morgan & Claypool Publishers, 2013.

- [13] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet. Audio information retrieval using semantic similarity. In *Proceedings IEEE Int. Conf. on Acoustics, Speech and Signal Processing. ICASSP 2007*, volume 2, 2007.
- [14] D. A. E. Beasley, E. P. Benson, S. M. Welch, L. S. Reid, and T. A. Mousseau. The use of Citizen Scientists to Record and Map 13-Year Periodical Cicadas (Hemiptera: Cicadidae: Magicicada) in South Carolina. *Florida Entomologist*, 95(2):489–491, 2012.
- [15] 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.
- [16] BirdLife International. Digital Distribution Maps of the Birds of the Western Hemisphere, version 5.0. BirdLife International and NatureServe, 2012.
- [17] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 4(2):1–22, 2009.
- [18] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009.
- [19] K. Bollacker, R. Cook, and P. Tufts. A platform for scalable, collaborative, structured information integration. In *International Workshop on Information Integration on the Web (VII IIWeb)*, 2007.
- [20] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM, 2008.
- [21] R. Bonney. Citizen science at the Cornell Lab of Ornithology. *Exemplary Science in Informal Education Settings: Standards-based Success Stories*, pages 213–229, 2007.
- [22] K. A. V. Borges, A. H. F. 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.
- [23] 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.



- [24] C. R. Buchanan. Semantic-based audio recognition and retrieval. Master's thesis, School of Informatics, University of Edinburgh, United Kingdom, 2005.
- [25] U. Caramaschi. Notes on the taxonomic status of *Elachistocleis ovalis* (schneider, 1799) and description of five new species of *Elachistocleis* Parker, 1927 (amphibia, anura, microhylidae). *Boletim do Museu Nacional. Nova Serie, Zoologia*, 527:1–30, 2010.
- [26] A. D. Chapman. Principles of data quality. *Report for the Global Biodiversity Information Facility. Copenhagen, Denmark.*, pages 1–58, 2005.
- [27] Cicada Project. <http://project.wnyc.org/cicadas/> (Accessed on: 06/2014).
- [28] M. Cobos and J. J. Lopez. Listen up - the present and future of audio signal processing. *IEEE Potentials*, 29(4):40–44, jul-aug 2010.
- [29] Cornell. The Cornell Lab of Ornithology. <http://www.allaboutbirds.org> (Accessed on 06/2014).
- [30] P. Cornillon, J. Gallagher, and T. Sgouros. Opendap: Accessing data in a distributed, heterogeneous environment. *Data Science Journal*, 2(0):164–174, 2003.
- [31] D. C. Cugler and C. B. Medeiros. Adaptive Acquisition of VGI to Fill out Gaps in Biological Observation Metadata (under review). In *Proceedings of the 22nd ACM International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, 2014.
- [32] D. C. Cugler, C. B. Medeiros, S. Shekhar, and L. F. Toledo. A geographical approach for metadata quality improvement in biological observation databases. In *Proceedings of the 9th IEEE International Conference on e-Science*, pages 212–220, 2013.
- [33] D. C. Cugler, C. B. Medeiros, and L. F. Toledo. Managing animal sounds - some challenges and research directions. In *Proceedings V Brazilian eScience Workshop - XXXI Brazilian Computer Society Conference*, July 2011.
- [34] 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*, 25(16):2310–2326, June 2012.
- [35] J. K. Eischeid, P. A. Pasteris, H. F. Diaz, M. S. Plantico, and N. J. Lott. Creating a serially complete, national daily time series of temperature and precipitation for the Western United States. *Journal of Applied Meteorology*, 39(9):1580–1591, 2000.

- [36] 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.
- [37] C. Fellbaum. *WordNet: An electronic lexical database*. The MIT press, 1998.
- [38] N. Fletcher. Animal bioacoustics. *Springer Handbook of Acoustics, ISBN 978-0-387-30446-5. Springer-Verlag New York, 2007, p. 785*, 1:785, 2007.
- [39] FNJV. Online animal sound collection - Fonoteca Neotropical Jacques Viellard. <http://proj.lis.ic.unicamp.br/fnjv> (Accessed on 06/2014).
- [40] Freebase. <http://www.freebase.com/>.
- [41] 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.
- [42] J. Futrelle, J. Gaynor, J. Plutchak, J. D. Myers, R. E. McGrath, P. Bajcsy, J. Kastner, K. Kotwani, J. S. Lee, L. Marini, R. Kooper, T. McLaren, and Y. Liu. Semantic middleware for e-science knowledge spaces. *Concurrency and Computation: Practice and Experience*, 23(17):2107–2117, 2011.
- [43] H. C. Gerhardt and F. Huber. *Acoustic communication in insects and anurans: common problems and diverse solutions*. University of Chicago Press, 2002.
- [44] M. F. V. Gomez and P. Willems. Filling gaps and disaccumulation of precipitation data for rainfall-runoff model. *Water observation and information systems for decision support*, 2:1–9, 2010.
- [45] M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007.
- [46] P. F. Gorder. Not just for the birds: archiving massive data sets. *Computing in Science Engineering*, 8(3):3–7, may-june 2006.
- [47] S. Gunasekaran and K. Revathy. Content-based classification and retrieval of wild animal sounds using feature selection algorithm. *International Conference on Machine Learning and Computing*, pages 272–275, 2010.
- [48] R. P. Guralnick, J. Wieczorek, R. Beaman, R. J. Hijmans, and the BioGeomancer Working Group. Biogeomancer: Automated georeferencing to map the world’s biodiversity data. *PLoS Biol*, 4(11):e381, November 2006.

- [49] C. F. B. Haddad. Anuran communication (in portuguese). *Anais de Etologia XIII. 1 ed. Pirassununga, SP, Brasil: Sociedade Brasileira de Etologia, v. 1*, 13:116–132, 1995.
- [50] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18, 2008.
- [51] C. Hartung, A. Lerer, Y. Anokwa, C. Tseng, W. Brunette, and G. Borriello. Open data kit: Tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 18. ACM, 2010.
- [52] C. Heipke. Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):550–557, 2010.
- [53] A. J. G. Hey, S. Tansley, and K. M. Tolle. *The fourth paradigm: data-intensive scientific discovery*. Microsoft research Redmond, WA, 2009.
- [54] W. R. Heyer and Y. R. Reid. Does advertisement call variation coincide with genetic variation in the genetically diverse frog taxon currently known as *Leptodactylus fuscus* (amphibia: Leptodactylidae)? *Anais da Academia Brasileira de Ciências*, 75(1):39–54, 2003.
- [55] L. L. Hill. *Georeferencing: The geographic associations of information*. MIT Press, 2009.
- [56] IBGE. Brazilian institute of geography and statistics. <http://www.ibge.gov.br> (Accessed on 05/2013).
- [57] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.
- [58] INMET. Instituto nacional de meteorologia. <http://www.inmet.gov.br/> (Accessed on 07/2014).
- [59] A. Irwin. *Citizen science: A Study of People, Expertise, and Sustainable Development*, volume 136. Routledge Londres, 1995.
- [60] IUCN International Union for Conservation of Nature. IUCN Red List of Threatened Species, 2012.
- [61] Java Script Object Notation JSON. <http://www.json.org>.

- [62] S. Kim, J. Mankoff, and E. Paulos. Sensr: evaluating a flexible framework for authoring mobile data-collection tools for citizen science. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1453–1462. ACM, 2013.
- [63] S. Kim, C. Robson, T. Zimmerman, J. Pierce, and E. M. Haber. Creek watch: pairing usefulness and usability for successful citizen science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2125–2134. ACM, 2011.
- [64] A. Kotsifakos, P. Papapetrou, J. Hollmén, and D. Gunopulos. A subsequence matching with gaps-range-tolerances framework: A query-by-humming application. *Proceedings of the VLDB Endowment*, 4(11):761 – 771, 2011.
- [65] R. K. Kumar and R. M. Chadrsekaran. Attribute correction–data cleaning using association rule and clustering methods. *Intl. Jrnl. of Data Mining & Knowledge Management Process*, 1(2):22–32, 2011.
- [66] A. Kurniawan. *Getting Started with Matlab Simulink and Raspberry Pi*. PE Press, 2013.
- [67] Kwantlen Polytechnic University. Plant Database. <http://plantdatabase.kwantlen.ca/> (Accessed on 06/2014).
- [68] 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.
- [69] T. B. Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.
- [70] B. D. Longueville, R. S. Smith, and G. Luraschi. OMG, from here, I can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 73–80. ACM, 2009.
- [71] 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.
- [72] J. E. G. Malaverri, M. S. Mota, and C. B. Medeiros. Estimating the quality of data using provenance: a case study in eScience. In *19th Americas Conference on Information Systems (AMCIS)*, 2013.

- [73] J. G. Malaverri, A. Santanchè, and C. B. Medeiros. A Provenance-based Approach to Evaluate Data Quality in eScience. *Int. J. Metadata, Semantics and Ontologies*, 2013.
- [74] J. G. Malaverri, B. Vilar, and C. B. Medeiros. A tool based on web services to query biodiversity information. In *5th International Conference on Web Information Systems and Technologies (Webist)*, pages 305–310, March 2009.
- [75] MeadoWatch. Citizen science at the Mount Rainier National Park to understand the biological impacts of climate change. <http://meadowatch.weebly.com> (Accessed on 06/2014).
- [76] B. Mechtley, G. Wichern, H. Thornburg, and A. Spanias. Combining semantic, social, and acoustic similarity for retrieval of environmental sounds. In *Proceedings IEEE Int. Conf. on Acoustics, Speech and Signal Processing. ICASSP 2010*, pages 2402–2405, march 2010.
- [77] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder. Discrimination and retrieval of animal sounds. In *Proceedings 12th IEEE International Multi-Media Modelling Conference*, 2006.
- [78] M. S. Mota, J. S. C. Longo, D. C. Cugler, and C. B. Medeiros. Using linked data to extract geo-knowledge. In *XII Brazilian Symposium on GeoInfomatics (GeoInfo)*, pages 111–116, Nov 2011.
- [79] Motorola. Shazam music recognition software. Available: <http://www.shazam.com/music/web/pages/shazamid.html> (Accessed on 11/2011).
- [80] G. Newman, J. Graham, A. Crall, and M. Laituri. The art and science of multi-scale citizen science support. *Ecological Informatics*, 6(3):217–227, 2011.
- [81] E. Posthumus and T. Crimmins. Nature’s notebook: A tool for education and research. *Bulletin of the Ecological Society of America*, 92(2):185–187, 2011.
- [82] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000.
- [83] R. Ranft. Natural sound archives: past, present and future. *Anais da Academia Brasileira de Ciências*, 76(2):455–465, 2004.
- [84] M. Rodell, P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, CJ Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, et al. The global land data

- assimilation system. *Bulletin of the American Meteorological Society*, 85(3):381–394, 2004.
- [85] D. Schorlemmer, A. Wyss, S. Maraini, S. Wiemer, and M. Baer. Quakeml: An xml schema for seismology. *Orfeus Newsletter*, 6(2):9, 2004.
- [86] S. A. Sheppard. wq: A modular framework for collecting, storing, and utilizing experiential VGI. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 62–69. ACM, 2012.
- [87] W. T. Silva. Development of a system to estimation of welfare from vocalization data of pigs (in portuguese). Master’s thesis, Univesity of Campinas - School of Agricultural Engeneering, 2008.
- [88] G. Solís, X. Eekhout, and R. Márquez. Fonoteca zoologica ([www.fonozoo.com](http://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.
- [89] R. B. Sousa, D. C. Cugler, J. E. G. Malaverri, and C. B. Medeiros. A provenance-based approach to manage long term preservation of scientific data. In *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*, pages 126–133. IEEE, 2014.
- [90] L. F. Toledo and C. F. B. Haddad. Acoustic repertoire and calling behavior of *Scinax fuscomarginatus* (anura, hylidae). *Journal of Herpetology*, 39(3):455–464, 2005.
- [91] L. F. Toledo and C. F. B. Haddad. Reproductive biology of *Scinax fuscomarginatus* (anura, hylidae) in south-eastern Brazil. *Journal of Natural History*, 39(32):3029–3037, 2005.
- [92] TV Cultura. Jornal da Cultura. <http://tvcultura.cmais.com.br/jornaldacultura/jornal-da-cultura-02-06-12-bl-02-mpeg2video-1> (Accessed: 12/2013), 02 June 2012.
- [93] J. C. Wang, H. P. Lee, J. F. Wang, and C. B. Lin. Robust environmental sound recognition for home automation. *Automation Science and Engineering, IEEE Transactions on*, 5(1):25–31, jan 2008.
- [94] K. D. Wells. *The ecology & behavior of amphibians*. University of Chicago Press, 2007.

- [95] G. Whitelaw, H. Vaughan, B. Craig, and D. Atkinson. Establishing the canadian community monitoring network. *Environmental monitoring and assessment*, 88(1-3):409–418, 2003.
- [96] M. Wick. Geonames. <http://www.geonames.org/> (Accessed on 12/2012).
- [97] J. Wimmer, M. Towsey, B. Planitz, P. Roe, and I. Williamson. Scaling acoustic data analysis through collaboration and automation. In *Proceedings Sixth IEEE International Conference on e-Science*, pages 308–315, 2010.
- [98] X. Xiao, P. Dorovskoy, C. Biradar, and E. Bridge. A library of georeferenced photos from the field. *Eos, Transactions American Geophysical Union*, 92(49):453–454, 2011.
- [99] 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.