UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

**Using a provenance model and spatiotemporal information to
integrate heterogeneous biodiversity semantic data**

Flor Karina Mamani Amanqui

Tese de Doutorado do Programa de Pós-Graduação em Ciências de
Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Flor Karina Mamani Amanqui

Using a provenance model and spatiotemporal information to integrate heterogeneous biodiversity semantic data

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Dilvan de Abreu Moreira

USP – São Carlos
December 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

M263u Mamani Amanqui, Flor Karina
Using a provenance model and spatiotemporal
information to integrate heterogeneous biodiversity
semantic data / Flor Karina Mamani Amanqui;
orientador Dilvan Moreira. -- São Carlos, 2017.
126 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2017.

1. Semantic Web. 2. Linked Open Data. 3.
Provenance. 4. Biodiversity. I. Moreira, Dilvan,
orient. II. Título.

Flor Karina Mamani Amanqui

Usando um modelo de proveniência e informações
espaço-temporais para integrar dados semânticos
heterogêneos sobre biodiversidade

Tese apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Doutora em Ciências – Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientador: Prof. Dr. Dilvan de Abreu Moreira

USP – São Carlos
Dezembro de 2017

This thesis is dedicated to my mother Florencia, my siblings Josimar and Stefany and my best friend and partner Evan for all of their continued love and support.

ACKNOWLEDGEMENTS

I would like to thank all the people who contributed in some way to the work described in this thesis. First and foremost, I want to thank my academic advisor, Professor Dilvan Moreira, for accepting me into his group. He provided a friendly and cooperative atmosphere at work and provided useful feedback and insightful comments for my work. Additionally, I would like to thank my committee members Professora Renata Pontin, Professor José Laurindo Campos, and Professora Karine Reis for their interest in my work.

First of all, I would like to thank the Institute of Mathematical and Computer Science (ICMC) at the University of Sao Paulo (USP) for providing the resources and facilities that I used for my research. I want to thank the various members of the Interactive Web and Multimedia Systems research group (Intermedia) with whom I had the opportunity to work. They supported me during my time in Brazil.

I also had the opportunity to work with Professor Erik Mannens at the University of Ghent from August 2015 until June 2016. I would like to thank Erik for his unconditionally support on my work. I would like to acknowledge the researchers from the Data Science Laboratory for their feedback and collaboration on my research.

I would like to acknowledge the Laboratory of Molecular Biodiversity and Conservation of the Federal University of Sao Carlos for their help with discovering issues and providing useful suggestions for this thesis.

There were many friends and family members who supported me during my PhD. First and foremost, I would like to thank my Mom Florencia, my siblings Josimar and Stefany, my boyfriend Evan, my syster in law Carolina, my niece Miranda, for their constant love and support. My friends from Perú, Brazil and Belgium. I am lucky to have met friends from different countries. Thank you to all of them for their support.

Finally, I wrote this thesis in different countries: Brazil, Peru, Belgium, United States, France, Switzerland. My research took me around the world attending conferences, and I want to thank the support from the National Innovation Program for Competitiveness and Productivity (Innóvate Perú), the Erasmus Mundus Program from the Europe Union, the Ghent University, iMinds, the IWT-Flanders, the FWO- Flanders. Without their support it would not be possible.

“Satisfaction lies in the effort, not in the attainment, full effort is full victory.”
“A nossa recompensa está no esforço, não no resultado. Um esforço total é uma vitória completa.”
“Nuestra recompensa se encuentra en el esfuerzo y no en el resultado. Un esfuerzo total es una victoria completa.”
(Mahatma Gandhi)

ABSTRACT

AMANQUI, F. K. **Using a provenance model and spatiotemporal information to integrate heterogeneous biodiversity semantic data.** 2017. 126 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

In the last few years, the Web of data is being rapidly populated with biodiversity data. However, when researchers need to retrieve, integrate, and visualize these data, they need to rely on semi-manual approaches. That is due to the fact that biodiversity repositories, such as GBIF, offer data as just strings in CSV format spreadsheets. There is no machine readable metadata that could add meaning (semantics) to data. Without this metadata, automatic solutions are impossible and labor intensive semi-manual approaches for data integration and visualization are unavoidable. To reduce this problem, we present a novel architecture, called STBioData, to automatically link spatiotemporal biodiversity data, from heterogeneous data sources, to enable easier searching, visualization and downloading of relevant data. It supports the generation of interactive maps and mapping between biodiversity data and ontologies describing them (such as Darwin Core, DBpedia, GeoSPARQL, Time and PROV-O). A new biodiversity provenance model (BioProv), extending the W3C PROV Data Model, was proposed. BioProv enables applications that deal with biodiversity data to incorporate provenance data in their information. A web based prototype, based on this architecture, was implemented. It supports biodiversity domain experts in tasks, such as identifying a species conservation status, by automating most of the necessary tasks. It uses collection data, from important Brazilian biodiversity research institutions, and species geographic distributions and conservation status, from the IUCN Red List of Threatened Species. These data are converted to linked data, enriched and saved as RDF Triples. Users can access the system, using a web interface, and search for collection and species distribution records based on species names, time ranges and geographic location. After a data set is recovered, it can be displayed in an interactive map. The records contents are also shown (including provenance data) together with links to the original records at GBIF and IUCN. Users can export datasets, as a CSV or RDF file, or get a print out in PDF (including the visualizations). Choosing different time ranges, users can, for instance, verify the evolution of a species distribution. The STBioData prototype was tested using use cases. For the tests, 46,211 collection records, from SpeciesLink, and 38,589 conservation status records (including maps), from IUCN, for marine mammal were converted to 2,233,782. RDF triples and linked using well known ontologies. 90% of biodiversity experts, using the tool to determine conservation status, were able to find information about dolphin species, with a satisfactory recovery time, and were able to understand the interactive map. In an information retrieval experiment, when compared with SpeciesLink keyword based search, the prototype's semantic based search performed, on average, 24% better in precision and 22% in recall tests. And that does not takes into account

cases where only the prototype returned search results. These results demonstrate the value of having public available linked biodiversity data with semantics.

Keywords: Semantic Web, Linked Open Data, Provenance, Biodiversity.

RESUMO

AMANQUI, F. K. **Usando um modelo de proveniência e informações espaço-temporais para integrar dados semânticos heterogêneos sobre biodiversidade**. 2017. 126 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Nos últimos anos, a Web de dados está sendo rapidamente preenchida com dados de biodiversidade. No entanto, quando pesquisadores precisam recuperar, integrar e visualizar esses dados, eles precisam confiar em abordagens semi-manuais. Isso ocorre devido ao fato de que repositórios sobre biodiversidade, como GBIF, oferecem dados como cadeias de caracteres em planilhas no formato CSV. Não há nenhum metadado legível por máquinas que poderia acrescentar significado (semântico) aos dados. Sem os metadados, soluções automáticas são impossíveis, sendo necessário para visualização e integração dos dados, a utilização de abordagens semi-manuais. Para reduzir esse problema, apresentamos uma arquitetura chamada STBioData. Com ela é possível vincular automaticamente dados de biodiversidade, com informações espaço-temporais provenientes de fontes heterogêneas, tornando mais fácil a pesquisa, visualização e download dos dados relevantes. Ele suporta a geração de mapas interativos e o mapeamento entre dados de biodiversidade e ontologias que os descrevem (como Darwin Core, DBpedia, GeoSPARQL, Time e PROV-O). Foi proposto um novo modelo de proveniência para biodiversidade (BioProv), que estende o modelo de dados PROV W3C. BioProv permite que aplicativos que lidam com dados de biodiversidade incorporem os dados de proveniência em suas informações. Foi implementado um protótipo Web baseado nesta arquitetura. Ele oferece suporte aos especialistas do domínio de biodiversidade em tarefas como, identificação do status de conservação da espécie, além de automatizar a maioria das tarefas necessária. Foi utilizado coleções de dados de importantes pesquisas brasileiras sobre biodiversidade, juntamente com dados de distribuição geográfica das espécies e seu estado de conservação, provenientes da lista de espécies ameaçadas da IUCN (Red List). Esses dados são convertidos em dados conectados, enriquecidos e salvados como triplas RDF. Os usuários podem acessar o sistema, usando uma interface web que permite procurar, utilizando os nomes das espécies, intervalos de tempo e localização geográfica. Os dados recuperados podem ser visualizados no mapa interativo. O conteúdo de registros também é mostrado (incluindo dados de proveniência), juntamente com links para os registros originais no GBIF e IUCN. Os usuários podem exportar o conjunto de dados, como um arquivo CSV ou RDF, ou salvar em PDF (incluindo as visualizações). Escolhendo diferentes intervalos de tempo, os usuários podem por exemplo, verificar a evolução da distribuição das espécies. O protótipo STBioData foi testado usando casos de uso. Para esses testes, 46.211 registros de coleção do SpeciesLink e 38.589 registros de estado de conservação da IUCN (incluindo mapas), sobre mamíferos marinhos, foram convertidos em 2.233.782 triplas RDF. Essas triplas reutilizam

ontologias representativas da área . 90% dos especialistas em biodiversidade, usaram a ferramenta para determinar o estado de conservação, eles foram capaz de encontrar as informações sobre determinada espécie de golfinho, com um tempo de recuperação satisfatório e também foram capaz de entender o mapa interativo gerado. Em um experimento sobre recuperação de informações, quando comparado com o sistema de busca por palavra-chave utilizado pela base SpeciesLink, a busca semântica realizada pelo protótipo STBioData, em média, é 24% melhor em testes de precisão e 22% melhor em testes de revocação. Não são considerados os casos onde o protótipo somente retornou o resultado da busca. Esses resultados demonstram o valor de ter dados conectados sobre biodiversidade disponíveis publicamente em um formato semântico.

Palavras-chave: Web semântica, Dados abertos vinculados, Proveniência, Biodiversidade.

LIST OF FIGURES

Figure 1 – Indication of the range and scale of the Web of Data originating from the LOD project.	34
Figure 2 – RDF example, the butterfly (S) was collected in the stage of life (P) larva (O).	38
Figure 3 – The architecture for a semantic search enabled biodiversity data repository.	43
Figure 4 – Web application for searching biodiversity information.	48
Figure 5 – Precision, recall and F1 experimental results.	51
Figure 6 – Areas where plant samples of the <i>phylum Tracheophyta</i> were collected in the state of Amazonas (Brazil).	53
Figure 7 – Spatial functions of GeoSPARQL.	58
Figure 8 – Examples of geometries represented in WKT.	59
Figure 9 – Example of provenance in biodiversity data.	61
Figure 10 – Architecture of publishing linked biodiversity data.	64
Figure 11 – Provenance model for biodiversity data.	65
Figure 12 – Example of the process to identify the conservation status of species and plot it in a map.	75
Figure 13 – Linked biodiversity data architecture.	80
Figure 14 – Example of BioProv provenance data.	84
Figure 15 – The STBioData Web interface supports searches based on species names and date range (a). It allows the visualization of collection datasets (d) in an interactive map (b), visualization of provenance information (c), download of datasets in RDF or CSV formats (e), online dataset statistics (f) and a SPARQL End Point (g).	85
Figure 16 – Statistics of biodiversity datasets stored in the STBioData prototype.	86
Figure 17 – Web interface of integrating existing data to determine the conservation status.	90
Figure 18 – The most common Keyword search tools used by the specialist in biodiversity to search occurrences of species.	93
Figure 19 – Precision and recall experimental results.	94
Figure 20 – Averaged 11-point precision/recall graph across 20 queries for a representative STBioData and SpeciesLink search system.	95
Figure 21 – STBioData architecture extended to link geospatial data.	102

Figure 22 – STBioData Web Interface extended to Link Geospatial Data. The Web interface supports search process within a range of date, GeoSPARQL functions (within and intersects), visualization of the linked biodiversity datasets in an interactive map, visualization of the provenance information, download the datasets in RDF or CSV formats and online dataset metrics (statistics). . . .	103
Figure 23 – Visualization of the provenance information in STBioData.	104
Figure 24 – Occurrences in South America.	105
Figure 25 – Occurrences organized by year collected.	105
Figure 26 – Map for a query in STBioData system. This map visualizes all the occurrence of specimens that intersects Amazon Forest and Pico da Neblina National Park. STBioData found 35 occurrence of specimens.	110
Figure 27 – Evaluation of STBioData and SpeciesLink keyword search system.	110
Figure 28 – Evaluation of the two systems using the amount of time in seconds to answer the 10 queries.	111

LIST OF ALGORITHMS

Algorithm 1 – Algorithm for mapping biodiversity data onto ontology terms	47
Algorithm 2 – Semantic search algorithm developed to compare ontology terms, derived from user input keywords, to the OntoBio graph in a triple store.	49
Algorithm 3 – Algorithm to compare retrieve, link and visualize linked biodiversity data.	87
Algorithm 4 – Query builder algorithm to compare retrieve, link and visualize linked biodiversity and geospatial data.	106

LIST OF SOURCE CODES

Source code 1 – Example of a SPARQL query	38
Source code 2 – Example of a triple map.	67
Source code 3 – Example of <i>Collecting</i> Activity.	69
Source code 4 – Example of spatiotemporal location.	69
Source code 5 – Example of <i>Cataloguing</i> Activity.	70
Source code 6 – SPARQL query used to obtain the provenance information of a specific dataset.	70
Source code 7 – SPARQL query used to retrieve collection records and conservation status of the <i>Inia geoffrensis</i> dolphin from the year 1987 to 2016.	91
Source code 8 – SPARQL query that use GeoSPARQL functions to recover occurrences of the species <i>Bacopa monnierioides</i> within <i>Pantanal</i> and <i>Cerrado Biome</i> between 1980 and 2000.	111

LIST OF TABLES

Table 1 – Summary of the features present in related works. OK indicates feature support, partial indicates partial support, and ‘–’ indicates no support. UO: Use Ontologies, UP: Use Provenance, STQ: SpatioTemporal Queries, OVM :Occurrence Visualization in a Map, SO: Search Occurrences, OS: Online Statistics, LOD: Linked Open Data.	79
---	----

LIST OF ABBREVIATIONS AND ACRONYMS

CoL	Catalogue of Life
DSL	Domain Specific Language
EMBRAPA	Brazilian Agricultural Research Corporation
ENVO	The Environment Ontology
GBIF	Global Biodiversity Information Facility
GeoSPARQL	A Geographic Query Language for RDF Data
GWT	Google Web Toolkit 2.6
INPA	National Institute for Amazonian Research
INPE	National Institute for Space Research
IRI	Internationalized Resource Identifier
LOD	Linking Open Data
MPEG	Emilio Gueldi Museum
NCBO	National Center for Biomedical Ontology
OBOE	Extensible Observation Ontology
OWL	Web Ontology Language
RDF	Resource Description Framework
RIA	Rich Internet Application
RML	RDF Mapping Language
SPA	Single Page Application
SPARQL	The Simple Protocol and RDF Query Language
URI	Uniform Resource Identifiers
URL	Uniform Resource Locator
VoMag	Vocabulary Management Task Group
W3C	World Wide Web Consortium
WKT	Well-Known Text

CONTENTS

1	INTRODUCTION	29
1.1	Restatement of the problem	29
1.2	Research questions and hypotheses	30
1.3	Main original contributions	31
1.4	Terminology and key concepts	32
1.4.1	<i>Biodiversity science</i>	32
1.4.2	<i>Linked Open Data</i>	33
1.4.3	<i>Semantic Web</i>	35
1.5	Outline	35
2	IMPROVING BIODIVERSITY DATA RETRIEVAL THROUGH SE- MANTIC SEARCH AND ONTOLOGIES	37
2.1	Key concepts	37
2.1.1	<i>RDF - Resource description framework</i>	37
2.1.2	<i>SPARQL - The simple protocol and RDF query language</i>	38
2.1.3	<i>OWL - Web ontology language</i>	39
2.2	Introduction	40
2.3	Related work	41
2.4	Semantic search architecture	43
2.4.1	<i>The biodiversity ontology</i>	44
2.4.2	<i>Species taxonomy</i>	45
2.4.3	<i>The mapping component</i>	46
2.4.4	<i>Web interface</i>	47
2.4.5	<i>The reformulator component</i>	47
2.5	Experiments	48
2.5.1	<i>Creation of triples from insect and fish collections</i>	50
2.5.2	<i>Semantic Search testing</i>	50
2.5.3	<i>Linking to amazon deforestation data</i>	52
2.6	Conclusions and future work	53
2.7	Remarks about the chapter	54
3	A MODEL OF PROVENANCE APPLIED TO BIODIVERSITY DATASETS	57
3.1	Key concepts	57

3.1.1	<i>GeoSPARQL</i>	58
3.1.2	<i>WKT</i>	58
3.1.3	<i>Provenance</i>	59
3.2	Introduction	60
3.3	Related work	62
3.4	Architecture for publishing linked biodiversity data	63
3.4.1	<i>Provenance model for biodiversity data (BioProv)</i>	64
3.4.2	<i>Mapping provenance and biodiversity data to RDF</i>	66
3.5	Use case	67
3.5.1	<i>Collecting activity</i>	68
3.5.2	<i>Cataloguing activity</i>	69
3.5.3	<i>Querying linked biodiversity data provenance</i>	70
3.6	Conclusion	71
3.7	Remarks about the chapter	71
4	LINKING BIODIVERSITY DATA USING SPATIOTEMPORAL INFORMATION	73
4.1	Introduction	73
4.2	Related work	76
4.3	The STBioData architecture	79
4.3.1	<i>Domain ontologies for biodiversity</i>	81
4.3.2	<i>The mapping component</i>	82
4.3.3	<i>STBioData web interface</i>	83
4.3.4	<i>The query reformulator component</i>	86
4.4	Experiments	87
4.4.1	<i>Creation of triples from heterogeneous sources</i>	88
4.4.2	<i>Integrating existing data to determine the conservation status</i>	89
4.4.3	<i>Use case testing: retrieving the conservation status of Inia geofrensis species from a date range</i>	92
4.4.4	<i>Biodiversity information retrieval experiment</i>	94
4.5	Conclusions and future work	95
4.6	Remarks about the chapter	96
5	LINKING BIODIVERSITY AND GEOSPATIAL DATA	99
5.1	Introduction	99
5.2	Extension of STBioData architecture	101
5.2.1	<i>Web interface</i>	102
5.2.2	<i>The query builder component</i>	104
5.3	Experiments	107
5.3.1	<i>GeoSpatial dataset</i>	107

5.3.2	<i>User Queries</i>	108
5.3.3	<i>Qualitative evaluation</i>	112
5.4	Related work	113
5.5	Conclusion and future work	114
5.6	Remarks about the chapter	114
6	CONCLUSION	115
6.1	Future work	116
6.2	International collaborations	117
6.3	Publications	118
	BIBLIOGRAPHY	119

INTRODUCTION

Biological diversity is essential to life sustainability on Earth ([MAGNUSSON, 2013](#)). The large amount of data generated by researchers in biodiversity has led to discussions about how to find the best way to organize these data and provide tools and environments that stimulate and facilitate the search for information. Currently, when using search tools for biodiversity data, experts specify their queries using one or more terms of interest. However, these terms may not match those that are part of the documents and, therefore, some relevant documents are not recovered ([AMANQUI *et al.*, 2014](#)) ([CARDOSO *et al.*, 2015](#)).

In Brazil, there is a network of Amazonian and extra-Amazonian institutions that are involved in studies of biodiversity. This network is integrated by important institutions, such as the National Institute for Amazonian Research (INPA) ([INPA.BR](#),), the National Institute for Space Research (INPE) ([INPE.BR](#),), the Global Biodiversity Information Facility (GBIF) ([GBIF.ORG](#),), the Emilio Gueldi Museum (MPEG) in Para([GOELDI.BR](#),), and the Brazilian Agricultural Research Corporation (EMBRAPA) ([EMBRAPA.BR](#),). These organizations collect and contribute large amounts of data about biodiversity.

One of the most frequent problems, reported by biodiversity researchers, is how to retrieve and integrate information simultaneously from the large number of data sources found on the various biodiversity databases. Typically, these users utilize the biodiversity data to visualize integrated information about the collected specimens ([SANTOS, 2003](#)) ([MAGNUSSON, 2013](#)).

1.1 Restatement of the problem

The problem is that a biodiversity domain expert may specify one or more terms (strings) for a data search and, due to the large amount of available data, get responses with too many results (not all relevant) ([MAGNUSSON, 2013](#)). He/she then has a lot of work sifting through the results for the desired information, because the results provided are very broad and may not

even contain the targeted data. This activity is not particularly well supported by biodiversity software tools based on keyword searching (the kind usually found in the Web) ([AMANQUI et al., 2014](#)). Even if a search is successful, it is the biodiversity specialist who must browse the selected documents to extract the information he/she is looking for. There is not much support for retrieving the actual information from the documents, a very time-consuming activity, and put it in a suitable format. Of course, there are tools that can retrieve texts, split them into parts, check the spelling, and count their words. But, when it comes to interpreting sentences and extracting useful information for biodiversity experts, the capabilities of current software are still very limited. It is simply very difficult to distinguish the meaning of the following query:

"Return all occurrences of records of insects that belong to the ant family (Formicidae) and have been found in an aquatic habitat in the Brazilian Amazon forest".

For instance, an SQL query, in a traditional database, would only succeed if records have the exact information (strings) searched in the query. In this case, a record of a *Paraponera clavata* specimen (bullet-ant) that was found in a *swamp* would not be returned. The strings *Paraponera clavata* and *swamp* are not in the query. Biodiversity experts also need more complex queries, e.g., requiring spatiotemporal query processing, such as deriving co-occurrences of species in a given space time frame. Such processing is seldom supported. Other queries involve biodiversity relations among species, e.g., farms within a protected area. Such relationships are not stored, and must be deduced by the expert after performing a sequence of queries and simulations.

1.2 Research questions and hypotheses

The main research question is:

- Research Question 1: How can we integrate heterogeneous biodiversity semantic data using their spatiotemporal information and provenance? We investigated the following main hypothesis:
 - Hypothesis 1: Representing biodiversity data as Linked Data will improve the integration of data from different and independent sources (if they share common ontology terms).

To answer the main question, we also needed to answer the following questions and hypothesis:

- Research Question 2: How can we improve the interoperability of biodiversity data?
 - Hypothesis 2: Representing biodiversity data, as Linked Data, will allow the use of more advanced and complex data queries, which were not possible before.
- Research Question 3: How can provenance be modeled for multiple biodiversity use cases?

- Hypothesis 3: Using the W3C PROV standard, we can model provenance in the biodiversity domain for multiple use cases.
- Research Question 4: How can we improve the location accuracy of biodiversity data?
 - Hypothesis 4: Formalizing spatiotemporal characteristics from biodiversity data will allow more accurate data location.

1.3 Main original contributions

The main original contributions of this thesis can be summarized by the following points:

- We have proposed a novel architecture, called STBioData, to automatically link spatiotemporal biodiversity data, from heterogeneous data sources, to enable easy searching, visualization and downloading of relevant data. It supports mapping between biodiversity data and the ontologies describing them and the generation of interactive maps (showing data). We were able to represent biodiversity data, as Linked Data, from different and independent sources (using CSV and ESRI formats). Biodiversity data was mapped to terms from relevant ontologies, such as Darwin Core, DBpedia, BioProv, Prov-O, Time and, GeoSPARQL, and stored using Semantic Web formats (such as the Resource Description Framework - RDF) and queried using Semantic Web tools (such as triple stores end points).
- We have proposed a conceptual model for provenance in biodiversity data (BioProv). This model is based in the W3C PROV Data Model. The PROV specification provides the concepts and supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments, such as the Web. The BioProv model was tested using a representative dataset, from important Brazilian biodiversity research institutions, and was able to model provenance in 3 different use cases.
- We built a web based prototype: the STBioData system¹ that implements the STBioData architecture. This system was used to automate the species conservation status identification task. For this task, it used collection data, from GBIF, and species geographic distributions and conservation status, from IUCN. These data were converted to linked data, enriched and saved as RDF Triples. Users can access the system, using a web interface, and search for collection and distribution records based on species names and time ranges. After a data set is recovered, it can be displayed in a geographic visualization. The records contents are also shown (including provenance data) together with links to the original records at GBIF and IUCN. When a user is happy with the data he got, he can export it as a CSV or

¹ <http://java.icmc.usp.br:2180/stbiodata/>

RDF file or get a print out in PDF. Choosing different time periods, users can verify the evolution of a species distribution.

- We have defined 4 use cases in biodiversity with features and scenarios to identify various user tasks: (i) Determine the conservation status of the *Inia geoffrensis* species using a date range; (ii) Create a report of species collected inside Brazilian biomes and natural parks; (iii) Classification of ecologically degraded areas; and (iv) Molecular identification of the alga *Cladophora delicatula*. Biodiversity specialists, from INPA and the Laboratory of Molecular Biodiversity and Conservation (from the Federal University of São Carlos - UFSCar), reviewed all 4 use cases and tested the STBioData system using the first two cases.
- We conducted quantitative and qualitative evaluations of the STBioData system to find out if it was able to help biodiversity experts retrieve and link biodiversity information. The system was evaluated by biodiversity experts from INPA and the Laboratory of Molecular Biodiversity and Conservation (UFSCar), and by Semantic Web specialists from the University of São Paulo and the Data Science Laboratory (from the University of Ghent - Belgium).
- The STBioData system was tested, by 16 biodiversity experts, using two use cases and a representative biodiversity dataset from INPA, the Botanical Institute (IBt/SP), EMBRAPA and the IUCN Red List of Threatened SpeciesTM(IUCN). 90% of users were able to find information about the species they were after. Search precision and recall were 24% and 22%, respectively, better than a popular biodiversity data site (SpeciesLink).
- We made freely available² linked biodiversity and geospatial information, generated by the STBioData system, for marine mammals. This information is encoded in 2,233,782 RDF triples and was linked to external sources using well known ontologies.
- This work generated 3 complete papers published in international conferences and 2 papers submitted to periodicals.

1.4 Terminology and key concepts

Before diving into this thesis contents, some essential concepts will be briefly introduced in this section. The following chapters will further expand these concepts whenever appropriate.

1.4.1 Biodiversity science

Biodiversity is a measure of the variety of organisms present in different ecosystems. Biodiversity data is about an assortment of different types of organisms that co-occur in time and

² <http://java.icmc.usp.br:2190/graphs>

space. Biodiversity data are collected in different parts of the world and published in different formats and patterns. Scientists rely on accessing and analyzing these diverse data, collected by communities of researchers, to have insights that can help issues of primary importance to science and society (GREEN *et al.*, 2005).

This proliferation of information, from different sources, means that the search for information could be met by a variety of available resources, which may store data about the same domains but have different characteristics. Therefore, the need for integration and analysis of data from these sources becomes more evident (ZIEGLER; DITTRICH, 2007). For biodiversity data to be analyzed and integrated efficiently, it is necessary to use new technologies to improve human and computer collaboration on the Web.

Semantic Web technologies were proposed to enable explicit declaration of knowledge embedded in data. Once semantic information is properly added to data, machines can use it to integrate information in an intelligent way (BERNERS-LEE; HENDLER; LASSILA, 2001). The main subject of this work is the use of semantic web technologies for the automatic integration of heterogeneous biodiversity data, using their semantic content and spatiotemporal information. Some of these technologies, used in this work, are presented in the next subsections.

1.4.2 *Linked Open Data*

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web (BERNERS-LEE, 2006). These practices can be summarized in the following four principles:

1. Things are identified with Uniform Resource Identifiers (URI). URI is a string of a standardized form that allows to uniquely identify resources (e.g., documents). Any resource on the Web is identified by a URI (KUHN; KAUPPINEN; JANOWICZ, 2014). A subset of URI is the Uniform Resource Locator (URL), which contains access mechanism and a (network) location of a document such as [<http://dbpedia.org/page/Beetle>](http://dbpedia.org/page/Beetle).
2. The used URIs are dereferenceable HTTP URIs. Dereferenced meaning that the computer can navigate to the URI using HTTP and request its representation intended for machines.
3. When a URI identifier is looked up, useful information is provided using standard formats.
4. Links to other URIs are provided, so that related information can be discovered.

The Linking Open Data (LOD)³ project is a grassroots community effort founded in January 2007 and supported by the W3C Semantic Web Group (CYGANIAK; JENTZSCH, 2014).

³ <http://lod-cloud.net/>

It refers to the use of Linked Data with open data. It has been adopted by an increasing number of data providers, leading to the creation of a global open data space containing billions of assertions - the Web of Data (HEATH; BIZER, 2011).

An indication of the range and scale of the Web of Data originating from the LOD project is provided in Figure 1. Each node in this cloud diagram represents a distinct data set published as Linked Data (CYGANIAK; JENTZSCH, 2014). The arcs, in Figure 1, indicate that links exist between items in the two connected data sets. Heavier arcs roughly correspond to a greater number of links between two data sets, while bidirectional arcs indicate outward links to the other exist in each data set.

Figure 1 – Indication of the range and scale of the Web of Data originating from the LOD project.

Source: Cyganiak and Jentzsch (2014).

Currently, the main producers of LOD are government authorities and scientists. Statistics

about RDF datasets are maintained by LODStats⁴, which lists nearly 9960 datasets with over 88 billion RDF triples (retrieved on 22 August 2017).

Linked data that has semantic information (metadata) attached to it can be understood by machines. That makes them more linkable and reusable. Semantic Web technologies aim to add semantics to this kind of data.

1.4.3 Semantic Web

The first definition of Semantic Web was coined by Berners-Lee et al. (BERNERS-LEE; HENDLER; LASSILA, 2001), they defined it like an extension of the current Web in which information is given well defined meaning, better enabling computers and people to work in cooperation. The key idea is to add semantics into the web content in order to make it easier to find and use for both humans and machines.

Other definition is from (BOLEY; TABET; WAGNER, 2001), who define the Semantic Web as the new-generation of the Web that tries to represent information in such a way that it can be used by machines, not just for display purposes, but for automation, integration and reuse across applications. These ideas and principles to enhance the Web are being put into practice under the guidance of the World Wide Web Consortium⁵.

A key concept, to realize the vision of the Semantic Web, is the ability to represent knowledge in structured collections of information and inference rules that computers can access and reason over. This means that the languages used are very flexible, and can support every desired scenario, in our case biodiversity domain. The need for expressive syntax and for unambiguous semantics lead to the development of ontologies.

An ontology is an explicit and formal specification of a conceptualization (GRUBER, 1995). Ontologies encode knowledge within a domain and also knowledge that spans domains. They include definitions of basic concepts in the domain and the relationships among them (DEVEDZIC, 2004). Life sciences scientists, including biodiversity experts, already use ontologies in their work. Therefore, it is easier for them understand their use to explain data.

1.5 Outline

This doctoral dissertation is an article thesis, a collection of research papers with an introduction and a conclusion chapters. In addition, each article has an introduction and conclusion sections placing the scope and results of the article in the thesis context. Five articles, co-authored by this thesis author, were organized in five chapters:

⁴ <http://stats.lod2.eu/>

⁵ <http://www.w3.org/>

- Chapter 1 describes the background and purpose of this thesis. This chapter is partly based on the following publication “Using Spatiotemporal Information to Integrate Heterogeneous Biodiversity Semantic Data” ([AMANQUI *et al.*, 2016a](#)).
- Chapter 2 describes how to improve biodiversity data retrieval through semantic search and ontologies. It is based on the publication “Improving biodiversity data retrieval through semantic search and ontologies” ([AMANQUI *et al.*, 2014](#))
- Chapter 3 presents a model of provenance applied to biodiversity datasets (BioProv). This chapter is based on the publication “A Model of Provenance Applied to Biodiversity Datasets” ([AMANQUI *et al.*, 2016b](#)).
- Chapter 4 describes the STBioData architecture and prototype system for Linked Biodiversity Data. It uses spatiotemporal information to integrate heterogeneous biodiversity data. This chapter is based on the following publications "STBioData: Linked Biodiversity Data using their Spatiotemporal Information, a Use Case about the Conservation Status of Amazon River Dolphins" ([AMANQUI; POTENCIANO; MOREIRA, 2017b](#)).
- Chapter 5 presents an extension of the STBioData architecture to answer more advances geospatial queries. This chapter is based on the following publication "An Architecture to Create Geospatial Queries for the Biodiversity Domain, a Use Case about Endangered Species in Brazilian Biomes" ([AMANQUI; POTENCIANO; MOREIRA, 2017a](#)).

Finally, in Chapter 6, we provide conclusions and discuss future works.

IMPROVING BIODIVERSITY DATA RETRIEVAL THROUGH SEMANTIC SEARCH AND ONTOLOGIES

In this chapter, the research question 2, "How can we improve the interoperability of biodiversity data?", is investigated. The article in this chapter, entitled "Improving Biodiversity Data Retrieval through Semantic Search and Ontologies", proposes an architecture that uses semantic search to query biodiversity data in semantic web formats. Semantic search aims to improve search accuracy by using ontologies to understand user objectives and the contextual meaning of terms, used in the search, to generate more relevant results. The article describes how the mechanism of mapping biodiversity data, in tabular format (CSV), to semantic data, in RDF, is designed and how the semantic search tool can find relevant information, based on ontologies. The architecture, presented in the paper, was incorporated into the STBioData architecture to implement semantic search. Before going into the article, we will present some important key concepts.

2.1 Key concepts

A key concept, to realize the vision of the Semantic Web, is the ability to represent knowledge in structured collections of information and inference rules that computers can access and reason over. This means that the languages used need to be flexible, support logical assertions, and be based on a standard recognized by the World Wide Web Consortium (W3C).

2.1.1 *RDF - Resource description framework*

Resource Description Framework (RDF) is a W3C standard language for data interchange on the Semantic Web ([LASSILA et al., 1998](#)). RDF extends the linking structure of the Web to

use URIs to name relationships between pairs of things. It is composed by triples, a set of two things related by a property. Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. Figure 2 shows a RDF triple. It is comprised of three pieces of information: Subject (S), Predicate (P), and Object (O). Where S and O are nodes and P is the property or aspect that relates the subject to the object. In Figure 2, the butterfly (S) was collected in the stage of life (P) larva (O). RDF offers the basic support for information interchange, but, for queering, another language is necessary.

Figure 2 – RDF example, the butterfly (S) was collected in the stage of life (P) larva (O).

Source: Elaborated by the author.

2.1.2 SPARQL - The simple protocol and RDF query language

The Simple Protocol and RDF Query Language (SPARQL) is a W3C standard language for querying, retrieving and manipulating RDF data (triples (PRUD'HOMMEAUX; SEABORNE, 2006)). On 21 March 2013, SPARQL 1.1 became an official W3C Recommendation. Source Code 1 shows a SPARQL query where a user wants to identify the stage of the life cycle of *Coleoptera*, an insect. The query shows all the information available for *Coleoptera*'s *:stageOfLife* property. Its anatomy is: declare prefix shortcuts (Line 1), define the query result clause (Line 2) and, finally, define the query pattern (Lines 3 to 5).

Source code 1 – Example of a SPARQL query

```

1: prefix :<http://examplebiodata.org/ns>
2: SELECT ?Species ?Result
3: WHERE {
4: ?Species :stageOfLife ?Result.
5: }
  
```

2.1.3 OWL - Web ontology language

RDF and SPARQL can handle representation and querying of basic LOD datasets, but they cannot handle ontology definitions. Ontologies encode knowledge within a domain and also knowledge that spans domains. They include definitions of basic domain concepts and the relationships (DEVEDZIC, 2004). The Semantic Web needs ontologies with a significant degree of structure. These need to specify descriptions for the following concepts:

- classes in the domains of interest;
- relationships that can exist among things;
- properties those things may have.

Ontologies are usually expressed in a logic-based language. It is then possible to build tools capable of doing automatic reasoning based on the facts in the ontology. More detailed ontologies can be created with the Web Ontology Language (OWL) (ALLEMANG; HENDLER, 2008), a W3C standard.

OWL¹ is a language derived from description logic that offers more constructs than RDF (ANTONIOU *et al.*, 2003). It is syntactically embedded into RDF and provides additional standardized vocabularies. OWL 2 is the most recent version of OWL. There are three different OWL 2 sublanguages or profiles. Each one offers a different level of expressivity (W3C, 2012).

OWL 2-EL is particularly suitable for applications employing ontologies that define very large numbers of classes and/or properties, but not instances. It guarantees execution in polynomial time (W3C, 2012).

OWL 2-QL is designed so that data, stored in standard relational database systems, could be queried, through an ontology, via a simple rewriting mechanism without any changes to the data (R; CALVANESE, 2012).

OWL 2-RL is aimed at applications that require scalable reasoning without sacrificing too much expressive power. This is achieved by defining a syntactic subset of OWL 2 that is amenable to implementation using rule-based technologies (W3C, 2012).

OWL have defined semantics that can be used for reasoning with ontologies and knowledge bases described using these languages.

The following article is organized as follows. Section 2.2 describes the paper's background. Section describes the related work. The proposed semantic search architecture is presented in Section 2.4. Section 2.5 discusses the experiments and the results for two experiments. In Section 2.6, we discuss the conclusion of our architecture for semantic search. Finally, Section 2.7 adds relates the article results to the thesis objectives.

¹ <http://www.w3.org/2004/OWL/>

This chapter is based on the following publication:

Improving biodiversity data retrieval through semantic search and ontologies, in: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, Vol. 1, 2014, pp. 274–281.

2.2 Introduction

Nowadays, the Web has become one of the main sources of biodiversity information. Biodiversity research institutions continually add new specimens and their related information to their biological collections and make this information available on the Web. These collections provide, among other things, detailed information about specimens distribution in space and time. Most specimen information indicates where the item was located, when it was collected and by whom ([SANTOS, 2003](#)). This information about location of specimen occurrences can be combined with other geo-referenced data for predictive distribution maps.

Even though the potential impact of using collection data with other geo-referenced data is enormous, the huge data volume of these collections, which continues to grow, is a difficult obstacle. Responding to the global interest in biodiversity conservation and sustainable development, several projects are under way to digitize important worldwide biodiversity collections. Some of these projects are: the Global Biodiversity Information Facility (GBIF), the Biodiversity Database Collection of the National Research Institute for the Amazon (INPA), Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA), Reference Center on Environmental Information (CRIA), and the New York Botanical Garden (NYGB). Many other projects exist with a mix of regional and/or specific aims. However, these projects do not have a standard or automatic way to represent their data and do not interoperate.

To find relevant data, from the huge amount of biodiversity information present on the Web, an efficient searching architecture is required. Search engines could be a solution to this problem of finding relevant biodiversity information from different sources. A search engine algorithm is based on trying to match keywords, from a user's keyword list, to strings from indexed records (e.g. from web pages) to generate a ranked list of search results.

Although search engines are very helpful in finding information on the Web (and get smarter all the time), they suffer from the fact that they do not know the meaning of the terms and expressions used in Web pages (or other kinds of records) and the relationships between them. In the biodiversity field, it is not different. The large quantity of data generated by research institutions is difficult to search. To overcome the search engine problems and to be able to retrieve relevant and meaningful information intelligently, the Semantic Web was proposed by ([BERNERS-LEE; HENDLER; LASSILA, 2001](#)).

The Semantic Web is considered the new-generation of the Web that tries to represent

information in such a way that it can be used by machines, not just for display purposes, but for automation, integration and reuse across applications (BOLEY; TABET; WAGNER, 2001). The key idea is to add semantics into the Web content in order to make it easier to find and use for both humans and machines.

The next generation of the Semantic Web promises to increase the performances and the relevance of search engines, by first attaching formal semantics to resources, and then exploiting this semantics during the search process (MIRON; GENSEL; VILLANOVA-OLIVER, 2009). According to (MANGOLD, 2007), the semantic search approach tries to augment and improve searches on a set of resources that are initially unknown to the user, by using ontologies and semantic annotations of these resources. Also, semantic search aims to improve search accuracy by understanding user objectives and the contextual meaning of the terms used in the search, as they appear in the searchable data, either on the Web or within a closed system, to generate more relevant results.

We present a new architecture that uses a semantic search system for biodiversity data and semantic web formats and tools to represent this data. It supports mapping between biodiversity data and the ontologies describing it. A prototype based on this architecture was implemented.

This prototype was tested using a set of representative data about biodiversity (206,000 records) from the National Institute for Amazonian Research (INPA) and Emilio Gueldi Museum in Pará (MPEG), two of the most important institutions doing research in biodiversity in the Amazon Forest. This data was downloaded from the SpeciesLink web site. SpeciesLink is a distributed information system that integrates primary data from biological collections from many research institutions from Brazil and abroad. It is also a popular online tool to search for biodiversity data. The test results showed a 28% improvement in precision and 25% in recall, when comparing our semantic search approach to keyword based search using the SpeciesLink search tool.

We also show easy data interoperability with other open data sources, which also use semantic web formats and ontology terms, through an example using deforestation data (from the National Institute of Space Research - INPE) to enrich collection data.

The remainder of this article proceeds as follows: Section 2.3 discusses related work. Section 2.4 shows the architecture for semantic search. Section 2.5 presents a synopsis of our experiments results and Section 2.6 concludes by summarizing our results and describing future works.

2.3 Related work

We studied a number of techniques for biodiversity information retrieval based in keyword based search and semantic search. The techniques for keyword based search basically

determine which collection records contain the keywords in the user query ([BAEZA-YATES; RIBEIRO-NETO, 1999](#)). A survey of the available literature indicates limitations in keyword based search techniques:

- According to ([SHARMA; DUHAN; SHARMA, 2010](#)), the search concentrates on the keyword matching of user query with indexed documents, while ignoring the semantic of the query. A term may have several synonyms that are not considered while returning the search results to the user due to their unavailability.
- Words used by users can have problems, such as synonym or words with many meanings, that are very difficult to solve. People often choose keywords subjectively, arbitrarily and lacking standardization. Information retrieval based on keywords (at the syntax level) focus on simply matching keywords, without the ability of knowledge representation, processing and understanding ([ZHAO *et al.*, 2010](#)).
- According to ([SANTOS; BAIÃO; TANAKA, 2011](#)), keyword-based search is not sufficient to capture the underlying semantics of user information needs, since it is content-oriented.

Even though keyword-based search have all this limitations, it is still the main and, in most cases, the only search tool available in major biodiversity repositories, such as:

- GBIF Data Portal ([GBIF.ORG](#),) is a service that provides access to millions of scientific data records about biodiversity that are being shared via the Global Biodiversity Information Facility (GBIF) network. In March 2014 there were 405,720,566 data records (352,593,699 georeferenced) accessible from this portal.
- SpeciesLink ([INPA, 2017](#)) is a distributed information system that integrates primary data from biological collections from diverse institutions, such as museums, herbaria and microbiological collections, from Brazil and abroad. It had, in March 2014, 326 collections and sub-collections 6,425,366 on-line records (2,719,146 georeferenced).

New search approaches have been proposed to overcome the terminology and meaning mismatch limitations in keyword based search. A number of techniques have been developed for using ontologies to retrieve relevant documents in response to a query. We list the ones we considered most related to the biodiversity field:

- In ([XIONG; HUANG; JIN, 2009](#)), a semantic search approach for geosciences is proposed in which a query agent is linguistically mapping lexicon vocabularies to concepts and relationships from geological ontologies.
- In ([BERKLEY *et al.*, 2009](#)), a semantic search system for ecology data is presented. It allows structured searches over user annotations using ontology terms. Authors have used the Extensible Observation Ontology (OBOE) for query expansion.

These systems use relational databases to store the biodiversity data and ontologies. The data being used, in each one, has to use the system's database schema forming a closed system. Most of the search techniques used require complex analysis, involving natural language processing, to discover the implicit context and semantics of query terms in relational databases (what is a limiting factor). Data stored in one system cannot be queried from another. Third part applications cannot easily query or share the data using, for instance, Linked Open Data (LOD) technologies.

2.4 Semantic search architecture

This section presents our semantic search architecture for biodiversity data. Figure 3 presents the architecture's overall schema. The development of this architecture was divided in

Figure 3 – The architecture for a semantic search enabled biodiversity data repository.

Source: Elaborated by the author.

three parts:

1. The Biodiversity Ontology, which play a central role in our semantic search architecture by providing a shared knowledge,

2. The Mapping Component, which maps the collection records to ontology entities,
3. The Web Interface, which process queries from either users, using a web interface, or machines, using a SPARQL Endpoint.

2.4.1 The biodiversity ontology

Among Semantic Web technologies, ontologies play a central role by providing a shared knowledge about the objects in the real world. They promote reusability and interoperability among different sources (ZHAO *et al.*, 2010). To deal with biodiversity data, we modified a biodiversity ontology (OntoBio) that is utilized to associate semantic meaning (terms) to data.

OntoBio was designed to conceptualize knowledge about biodiversity collection data. Originally created by INPA (ALBUQUERQUE; SANTOS; CASTRO, 2015), it is being jointly developed by it and us (at ICMC - University of São Paulo). Its main objective is to provide a clear and precise conceptualization of the information describing specimen's collections. Ontobio is divided into five sub-ontologies (Collection, Material Entity, Spatial Location, Ecosystem, and Environment), integrated by relationships between their concepts and axioms. OntoBio was modeled using the OntoUML language, as its formal language for conceptual modeling, allowing it to capture complex aspects of the biodiversity domain. The development of OntoBio is only being possible due to the help of the highly capable experts, from INPA, willing to contribute to the project. The complete ontology is presented in details in (ALBUQUERQUE; SANTOS; CASTRO, 2015).

One of the advantages of having data annotated using OntoBio concepts (for that matter, using any open ontology) is that it can be reused as Linked Open Data (LOD). LOD describes a method of publishing structured data so that it can be interlinked and become more useful (KAUPPINEN; ESPINDOLA, 2011). To better archive that, data annotated using OntoBio has to be easily interlinked with other data already available on the web (as part of the wider LOD community) through the use of as many shared concepts as possible. With that in mind, we rewrote the original version of OntoBio to reuse, whenever possible, terms from other public available ontologies to allow better "linkability" with data already annotated using them.

When reusing an element from another ontology, we copied its URI and any axioms related to it that we needed. Then, if necessary, we added new axioms to it. We added terms from the following public ontologies or controlled vocabularies (all available in the OWL or RDF languages):

- The Environment Ontology (ENVO) (BUTTIGIEG *et al.*, 2013), which provides a controlled, structured vocabulary that is designed to support the annotation of any organism or biological sample with environment descriptors. EnvO contains terms for biomes, environmental features, and environmental material. In OntoBio, we use it to describe biomes (ENVO:00000428) and other environmental features. Examples of biome terms are: boreal

moist forest biome, tropical rainforest biome, and oceanic pelagic zone biome. ENVO is available to view or download in the Bioportal public Web site. The BioPortal is a Web portal that provides access to a library of biomedical ontologies and terminologies via the National Center for Biomedical Ontology (NCBO) Web services.

- The Darwin Core Standard (BASKAUF; WEBB, 2016) includes a glossary of terms intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. In OntoBio, we use it to describe properties, elements, fields, columns, attributes and concepts.
- The Geonames Ontology (WICK, 2015), which makes it possible to add geospatial semantic information to the Word Wide Web. Over 8.3 million geonames toponyms now have a unique URL with a corresponding RDF web service. Other services describe the relation between toponyms. The GeoNames Ontology is available in OWL.
- The DBpedia Ontology (LEHMANN *et al.*, 2015), which is a community-curated ontology consisting of 320 classes which form a subsumption hierarchy and are described by 1,650 different properties. The ontology is maintained and extended by the community in the DBpedia Mappings Wiki. This community also creates mappings from Wikipedia information representation structures (info-boxes created by Wikipedia editors) to the DBpedia ontology. These mappings are used to automatically create instances of the ontology from Wikipedia information (in Wikipedia's many languages), which ensures a huge coverage of topics. OntoBio uses instances from the English (dbpedia.org) and Portuguese (pt.dbpedia.org) mappings of DBpedia. The Portuguese mapping is mainly used used to describe Brazilian cities².

The Protégé 4 ontology editor was used to write the new OntoBio ontology version in OWL 2 DL. This new version has a dereferenceable URI³, meaning that the URI can be used by tools, e.g. Protégé 4, to get the ontology automatically from the web (as an OWL file).

2.4.2 Species taxonomy

In addition to OntoBio, we need a biological taxonomy to classify species. A taxonomy is just an ontology where each class can have just one parent. Unfortunately, there is no standard taxonomy used by all biologists. Biodiversity experts from INPA recommended the use of the taxonomy created by the Catalogue of Life⁴, an online database of the world's known species of *animals, plants, fungi and micro-organisms* (with 1.35 million species). The Catalogue of Life taxonomy is not available for download as a separate file in RDFS or OWL. So we had to write a

² <http://pt.dbpedia.org/resource/Urucurituba>

³ <http://purl.org/biodiv/ontobio>

⁴ <http://www.catalogueoflife.org>

program to use the site web services description ⁵ to navigate through its taxonomic tree and write it as an OWL file.

2.4.3 The mapping component

The Mapping Component loads the domain ontologies, taxonomic information and the biodiversity data collection to transform them in a set of RDF triples. We used a small Domain Specific Language (DSL) to represent the mapping between rows of data tables into OntoBio classes and properties, to create RDF triples.

Data from INPA and MPEG, and from dozens of other biodiversity research institutions, is available in the SpeciesLink web site. SpeciesLink offers this data in csv text files using a format based on Darwin Core. We used the mapping component to convert all INPA's and MPEG's records for the Brazilian state of Amazonas from the SpeciesLink web site to RDF triples. This mapping is done offline and generates the triples that will be stored in the triple store (in our case, GraphDB) and queried during user searches.

GraphDB (GüTING, 1994) is a triple store with very good performance and also works with multiple RDF graphs (knowledge trees) at the same time and supports the SPARQL 1.1 query language and GeoSPARQL functions. It also provides a faceted browser user interface for querying the RDF data store.

This mapping is illustrated in the [Algorithm 1](#).

This algorithm is capable of:

1. Create an ontology individual (entity) representing each specimen in the collection.
2. Automatically link specimen name to taxon data using the Catalogue of Life (CoL)⁶ webservices. Each collection record receives a URI connecting it to a taxon id in the CoL website;
3. Automatically link geographic information to the DBpedia, the Wikipedia Linked Data version. For instance, the DBpedia URI for each city is added to the record of each specimen collected;
4. Convert strings representing dates in various formats to the proper equivalent XSD date type (used by RDF) and check for semantic errors, such as an animal species being declared as belonging to the plant kingdom.

After mapping the data from the repository into RDF triples, they are loaded into a triple store (in this case, GraphDB) and used by the Web Interface to do its SPARQL queries.

⁵ <http://webservice.catalogueoflife.org>

⁶ <http://www.catalogueoflife.org/>

Algorithm 1 – Algorithm for mapping biodiversity data onto ontology terms

```

1: Input: Collection data in csv format
2: Output: Owl file with ontology and triples
3: for rows in Collection do
4:   Get species name of specimen
5:   Use CoL webservices to link name to taxonomy id
6:   if species name not found then
7:     Try next taxonomic rank until CoL can find an id
8:   end if
9:   Create an OWL individual to represent specimen
10:  Assert that it belongs to taxon returned by CoL
11:  Add label with taxon name
12:  Get municipality of colection
13:  Use DBpedia webservices to find municipality and state URIs
14:  Add municipality and state URIs
15:  if latitude and longitude available then
16:    Convert them to proper format
17:    Add them to OWL individual
18:  end if
19:  Add OWL Literals for colection and classification dates
20:  Add other information, such as gender, institution, etc
21:  Check for semantic errors in individual
22: end for
23: Dump OWL file to triple store

```

2.4.4 Web interface

The Web Interface is responsible for the interaction between users and our semantic search engine. The search process begins with an initial keyword list entered by a user (biodiversity specialist) that represents his/her search intentions. When typing these keywords, a widget from BioPortal (called term-selection field) suggests new terms based on semantic expansion and dictionary similarity with terms from BioPortal hosted ontologies (OntoBio included). Users may or may not accept the suggested terms. The search results display consists of a vertical list of document titles and several lines from the records that fulfill the search criteria (family, genus, specie and other information). The user interface is shown in [Figure 4](#) The queries entered in the Web interface are processed by the query reformulation component.

2.4.5 The reformulator component

The Reformulator Component is done online and receives the search terms from the user, it finds all classes of ontologies to which the term belongs and expands the keyword list using an [Algorithm 2](#). The basic idea of this algorithm is to compare ontology terms, derived from user input keywords, to the OntoBio graph in the GraphDB triple store.

When a user submits a query, each word in the query is compared with the labels from

Figure 4 – Web application for searching biodiversity information.

Source: Elaborated by the author.

the triples in the GraphDB triple store (using SPARQL queries). This includes OntoBio terms and collection data. All relevant triples are found.

If a triple also has latitude and longitude data attached to it, these are also collected. The algorithm results, obtained using multiple SPARQL queries, are presented to the user in the Web Interface.

This algorithm was implemented in a prototype, a Rich Internet Application (RIA), using the Java technology. The server side used the Jena RDF framework, to reason about SPARQL queries, in addition to a GraphDB triple store. The Web Interface was implemented using Google Web Toolkit 2.6 (GWT) on the client side.

2.5 Experiments

In order to validate our architecture and guide the prototype tests, INPA's biodiversity experts were interviewed to categorize important information from INPA's and MPEG's data (e.g. genus, family, species, description of location, etc.). These interviews helped us to understand

Algorithm 2 – Semantic search algorithm developed to compare ontology terms, derived from user input keywords, to the OntoBio graph in a triple store.

```

1: Input: userQuery a set of words typed by user
2: Output: results a information list from the triples
3: Connect to Triple Store and OntoBio ontology
4: for word in userQuery do
5:   submit a SPARQL query with word as subject
6:   add return to results                                ▷ word ?predicate ?object
7:   submit a SPARQL query with word as predicate
8:   add return to results                                ▷ ?subject word ?object
9:   submit a SPARQL query with word as object
10:  add return to results                                ▷ ?subject ?predicate word
11: end for
12: for unique objects and subjects in results do
13:   submit a SPARQL query to find its location
14:   add return to results                                ▷ :longitude and :latitude
15: end for

```

more about their work and to form a common ground for discussions. Because this technology is so new, it is difficult for us and our partners at INPA to foresee all its possible uses. To help us, we defined use cases with features and scenarios to identify the various user tasks. One such use case is presented below:

- **USE CASE:** Classification of Ecologically Degraded Areas
- **USER:** Christine Smith, 32 years-old, biologist, NGO employee.
- **GOAL:** To determine if areas in the state of Pará, Brazil, are ecologically degraded based on the size of their deforested areas and species collected there.
- **MOTIVATION:** The presence or not of some species of plants and animals can serve as biological markers (bio indicators) that indicate the degree of conservation or degradation in a habitat.
- **TASKS**
 1. Find deforestation information: The Linked Brazilian Amazon Rain Forest Data SPARQL EndPoint divides the Brazilian forest in 25 km squares with deforestation information.
 2. Link the geographic information of collected specimens to their deforestation level.
 3. Use the information to plot maps using tools such as the R language (software environment for statistical computing and graphics).
- **NECESSARY TOOL FEATURES**

1. Specification of bio-marker species using the species name or any higher taxonomic level, like phylum, genus or family.
2. EndPoint SPARQL to answer queries from computer programs, such as the R script, to allow data integration with other repositories.
3. Flexible way to limit habitats of interest (one species can be a bio-marker in one habitat but not in another). For example, Christine can specify an aquatic habitat for an insect without having to specify if this habitat is a river, stream or lake.

After studying our use cases, we divide the experimental evaluation of our prototype in three parts: creation of triples using INPA's and MPEG's insect and fish collections, semantic search testing and linking with deforestation data.

2.5.1 Creation of triples from insect and fish collections

The first experiment demonstrated the mapping mechanism, using the OntoBio ontology and INPA's and MPEG's collections data, to obtain a set of triples (subject, object and predicate). We used the [Algorithm 1](#) to convert all INPA's and MPEG's records for the Brazilian state of *Amazonas* from the SpeciesLink web site (206,000 records) to RDF triples. This RDF data was stored in our GraphDBTriple Store and can be explored using SPARQL queries. The mapping program was able to treat defective input records, such records lacking fields. The experiment demonstrated that the mapping method works even when the data being used has defective records. To integrate the biodiversity data in RDF to the wider LOD data community on the web, we setup a SPARQL EndPoint⁷. Our EndPoint allows that third part programs query our knowledge base, via the SPARQL language, and reuse it in their applications.

2.5.2 Semantic Search testing

For the previous use case, biodiversity experts from INPA identified the information set (from the data converted to RDF triples) that each user needed for each task. They also created queries examples that should have returned this information. After we tested each query, the same experts judged which results were relevant and non relevant. In the following, we show examples of user tasks and possible query strings to get the relevant biodiversity information.

- User Tasks
 - **Differentiate scientific name:** the scientific name is what scientists sometimes call something to differentiate it from other things in the same order/family/genus. For example: *fish ocellatus*, *fish brasiliensis*, *Corydoras guianensis*, *steindachneri* *Crenicichla*, *Geophagus*.

⁷ <http://java.icmc.usp.br:2190/sparql>

- **Determine if group members breed with organisms outside their group:** a difficult task for biologists is to determine if a group of organisms is a separate species. For example: *Hypsiboas*, *Anolis*, *hoplias*, *camponotus jupiaba*.
- **Determine the geographic distribution of specimens:** a common task for biologists is the monitoring of wildlife population levels. For example: *nannostomus bryconops*, *Serrasalmus acestrorhynchus*.

Researchers from INPA and system analysts from USP participated in the experiments. A total of 28 distinct queries were done. It is important to remember that it takes a lot of effort, from the domain experts, to figure out which data should be returned in each query, given the size of the data set, and the relevance of the returned results. We tested and compared the result of two search systems: our semantic search engine and the keyword based search system used by the SpeciesLink site (this site uses a standard search system for biodiversity data), both using data from INPA and MPEG. The evaluation of the two search approaches, using the average precision, recall and F1 is shown in Figure 5. The F1 score can be interpreted as a weighted average of the precision and recall scores. A F1 score reaches its best value at 1 and worst at 0. Our semantic search architecture reached a highest precision around 28% (on average) better

Figure 5 – Precision, recall and F1 experimental results.

Source: Elaborated by the author.

than using the keyword search, in the scenarios described by the biodiversity experts.

The semantic search recall was 25% (on average) better than the keyword based search (from SpeciesLink). This results demonstrate that the values for precision and recall, for the semantic search, were significantly better.

In all use case scenarios, the semantic search obtained better F1 measure results (26% better on average than the keyword search), reflecting a good balance between the increased precision and recall obtained.

2.5.3 Linking to amazon deforestation data

Semantic or keyword based search can be useful to peruse data (to find out if it is useful for some application) or to find a small number of specific records. But when it comes to data analysis, computer programs are necessary. There are two ways to write these computer programs:

- Applications developed by programmers for use by biologists.
- Scripts written by biologists in some DSL.

Lets focus on the second case, it is simpler and does not need any programmers to work.

For this test, we downloaded all INPA's records (August 2013) for the state of Amazonas - Brazil (108,220 in total) and, using a script and a mapping (written in a small DSL), read all plant records (53,120 in total) and transformed them into RDF triples using the OntoBio ontology. These triples were loaded and made available in our GraphDB triple store. Finally we chose the "Linked Brazilian Amazon Rainforest" dataset ([KAUPPINEN; ESPINDOLA, 2011](#)) to extract information to connect to our dataset. This dataset consists of 8250 cells—each of size of 25 km by 25 km—capturing the observations of deforestation in the Brazilian Amazon Rainforest and a number of related and relevant variables. It is available in a SPARQL EndPoint⁸ and via dereferenceable URIs.

With the help of R, a DSL language popular among biodiversity researches, and its packages sparql and so we created a script to:

1. read, using a SPARQL query, all the cells belonging to the Amazonas state from the "Linked Brazilian Amazon Rainforest" dataset with their positions (a polygon) and deforestation percentage (2008 values).
2. read, using a SPARQL query, all occurrences containing plant samples belonging to the phylum Tracheophyta that had latitude and longitude information (19,887) from our INPA's dataset.
3. for each cell from step 1, which had a deforestation ratio smaller than 10%, count the number of collections made inside it.
4. plot the cells with colors representing the number of collections in each.

⁸ <http://spatial.linkedscience.org/sparql>

Figure 6 shows the generated map. The brown color shows areas with deforestation bigger than 10% (no specimens counted). The map shows a fairly amount of specimens coming from a small area. The black square has the biggest number of collections, 835, and is far greater than the next one, with 380. We separated it to get a better color distribution in the map.

The map shows that we can merge our data with other Linked Data resource to enhance its informative value. Using the results, users can know which collections come from areas there are still preserved and see their distribution in a map. One of the big advantages of having INPA's data in RDF is to be able to connect it to other sources. In Figure 6, we show areas where plant

Figure 6 – Areas where plant samples of the *phylum Tracheophyta* where collected in the state of Amazonas (Brazil).

Source: Elaborated by the author.

samples of the *phylum Tracheophyta* where collected in the state of Amazonas (Brazil). The color brown indicates areas with more than 10% deforestation (2008 data), samples from them were not counted. The distribution is quite concentrated in a few areas, the black spot represents the area with more collections (830).

2.6 Conclusions and future work

We presented a new architecture that uses a semantic search system to query biodiversity data and semantic web formats and tools to represent it. We implemented a prototype of it and, with it, showed that the architecture produces results with better precision, recall and F1-measure than a standard biodiversity search tool (SpeciesLink) using real biodiversity data from INPA and MPEG collections.

We demonstrated that, once collection data is in RDF format, it is easy to integrate it with data from different and independent data sources (if they share common ontology terms).

We also showed that it is possible, for a biodiversity expert, to write a DSL script himself to do this data integration and answer scientific questions (using R, a popular scripting language among biologists). Data integration and scripting are very important because, in many cases, after finding relevant data, it is almost impossible to manipulate it by hand given its size. So, experts have to write scripts to integrate and analyze data to get the answers to their scientific questions.

As future work, we plan to refine our use cases. We intend to reuse GeoSPARQL⁹ ontology terms to describe georeferenced data, such as shapes of municipalities, national parks and farms, to include geographical semantic information in queries. We also intend to extend our current implementation with more advanced structured search types, in partnership with INPA's researchers.

2.7 Remarks about the chapter

In this chapter, we responded research question 2: "How can we improve the interoperability of biodiversity data?". An architecture was proposed to convert data to semantic web formats and search biodiversity data. This chapter describes how the mechanism of mapping is designed so that the semantic search can find information, based on ontologies.

The article demonstrated that, once biodiversity collection data is in RDF format, it is easy to link it to other data sources to find new results (interoperability). The collection records, from INPA, were linked to the Linked Brazilian Amazon Rainforest datasets (already in RDF). Using the combined dataset a map was generated showing collections in areas with 10% or less deforestation. It took one SPARQL query to get the relevant map data (improved interoperability). The map showed that a fair amount of specimens collected comes from a small area.

Also, a semantic search prototype (using the proposed architecture) reached a precision 28% (on average) better than using keyword search (from SpeciesLink), in the scenarios described by the biodiversity experts. The semantic search recall was 25% (on average) better than the keyword based search. These experiments demonstrated that, once in semantic web format, we can merge our data with other Linked Data resources to enhance its informative value. Using the results, users can now know the preservation state of the areas where the collections came from and see this information in a map.

The possibility of creating queries that seek information based on relationships between different datasets offers many alternatives to semantic search systems, since the results of these queries are not based only on specific information (one data source). Users can find data that, in traditional systems, would not be considered by the query.

In the next chapter, we will extend the current architecture to support provenance infor-

⁹ <http://www.opengeospatial.org/standards/geosparql>

mation and more advanced structured queries and will describe how provenance can help model biodiversity data.

A MODEL OF PROVENANCE APPLIED TO BIODIVERSITY DATASETS

In this chapter, the research question 3, "How can provenance be modeled for multiple biodiversity use cases?", is investigated. The chapter's article, entitled "A Model of Provenance Applied to Biodiversity Datasets", proposes an extended provenance model applied to biodiversity datasets (BioProv). It describes how provenance can be modeled in the biodiversity domain. We will convert to RDF a representative dataset about biodiversity, from the Botanical Institute (IBt/SP), and add provenance information for the use case of Molecular Identification of *Cladophora delicatula* Alga. The BioProv model, presented in the paper, was incorporated into the STBioData architecture to provide provenance information. Before going into the article, we will present some key concepts for this chapter.

3.1 Key concepts

The use of ontologies and LOD provide a huge potential in various domains. In the case of biodiversity, data is very geospatial dependent. For instance, a collection is made in a geographic location, an species occurs in a certain area. So, in biodiversity, geospatial data is pervasive and, because of that, can be used for data integration from different sources.

Linked Geospatial Data has recently received attention as researchers and practitioners have started tapping the wealth of geospatial information, available in the archives of various national cartographic agencies, and making it available on the Web as Linked Data ([ABBAS; OJO, 2014](#)).

As a result, in the last few years, the Web of Data is being rapidly populated with geospatial information. As the real-world entities, represented in Linked Geospatial Data sets, evolve over time, the data sets themselves get updated and both the spatial and the temporal dimensions of data become significant for users. To describe Geospatial data, it is necessary to

use a new ontology in the Semantic Web: GeoSPARQL.

3.1.1 GeoSPARQL

A Geographic Query Language for RDF Data (GeoSPARQL) is a standard for representing and querying Linked Geospatial Data for the Semantic Web from the Open Geospatial Consortium ([CONSORTIUM, 2017](#)).

The definition of a small ontology based on OGC standards is intended to provide a standardized exchange basis for geospatial RDF data. It can support both qualitative and quantitative spatial reasoning and querying within the SPARQL language ([WIEGAND *et al.*, 2007](#)).

There are three main components of GeoSPARQL. The first component is the definition of a vocabulary to represent features, geometries and their relationships. The second is a set of domain specific spatial filter functions for use in SPARQL queries. The last component is a set of query transformation rules. Three main classes are defined in the GeoSPARQL ontology: *ogc:SpatialObject*, *ogc:Feature* and *ogc:Geometry*.

An example of GeoSPARQL spatial functions is showed in [Figure 7](#). For example, to determine whether two objects A and B overlap, we can use the *geo:overlaps* function.

Figure 7 – Spatial functions of GeoSPARQL.

Source: [Wiegand *et al.* \(2007\)](#).

The GeoSPARQL ontology works as an extension to the SPARQL language that allows queries based on spatial relations, such as being inside or to border, and the representation of coordinates in the form of points, lines and polygons. To represent geographical coordinates in GeoSPARQL, the OGC defines a pattern called Well-Known Text (WKT).

3.1.2 WKT

WKT is a text markup language for representing vector geometry objects (on a map), spatial reference systems of spatial objects and transformations between spatial reference systems ([OGC, 2017](#)).

Geometries in WKT are restricted to 0, 1 and 2 dimensional geometries that exist in the R2, R3 or R4 spaces. Geometries that exist in R2 consist of points with coordinates x and y, e.g., POINT(3 5) in WKT syntax.

Geometries that exist in R3 consist of points with coordinates x, y and z, or x, y and m where m is a measurement. For example, the point POINT(-3 -60 32) might be used to represent the temperature of the city of Manaus (measured in degrees Celsius); where -3 is the latitude of Manaus, -60 its longitude and 32 its temperature.

Geometries that exist in R4 have points with coordinates x, y, z, and m with similar semantics (KOUBARAKIS *et al.*, 2012). Some examples of geometries represented in WKT are shown in Figure 8.

Figure 8 – Examples of geometries represented in WKT.

Geometry type	WKT representation	Geometry
Point	Point(5 5)	.
LineString	LineString(5 5,28 7,44 14,47 35,40 40,20 30)
Polygon	Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5))
Polygon	Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5), (28 29,14.5 11,26.5 12,37.5 20,28 29))
MultiPoint	MultiPoint((5 5),(28 7),(44 14), (47 35),(40 40),(20 30))
Geometry Collection	GeometryCollection(Point(5 35), LineString(3 10,5 25,15 35,20 37,30 40), Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5), (28 29,14.5 11,26.5 12,37.5 20,28 29)))

Source: Koubarakis *et al.* (2012).

The interpretation of geometry coordinates depends on the coordinate reference system that is associated with that geometry. Note that, according to the WKT standard, the coordinate reference system that is associated to a geometry is never embedded in the object representation, but it is given separately using appropriate notation. In this work, the WKT representation will be used to represent geographic information in biodiversity data.

3.1.3 Provenance

Data provenance, also called data lineage, records sources as well as a set of processing steps applied to sources (CELINO, 2013). The new standard for provenance data is W3C PROV Data Model (MOREAU; Missier (Eds.); W3C Provenance Working Group, 2013b). This standard

was defined by the World Wide Web consortium and is based on an ontology, PROV-O expressed in OWL 2. The PROV-O specification (W3C, 2013) provides the concepts and supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web. This conforms to the principles of LOD, which encourages a set of best practices for publishing and connecting structured data on the Web.

The rest of this chapter is organized as follows. Section 3.2 describes the article background. Section 3.3 describes the related work. The provenance model for biodiversity datasets is presented in Section 3.4. Section 3.5 presents a use case, where we model all 217,829 plant records of the SpeciesLink web site using PROV-O. Section 3.6 concludes by summarizing the article results and describes future work. Finally, the remarks of the article are given in Section 3.7.

This chapter is based on the publication:

A Model of Provenance Applied to Biodiversity Datasets. 25th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2016). Paris, June 13-15, 2016.

3.2 Introduction

Biological diversity is essential to life on Earth and motivates many efforts to collect data about species (MAGNUSSON, 2013). That gives rise to large amounts of data. These data are collected in different places and published in different formats. Collecting data in the field is expensive, difficult, and sometimes dangerous. Not only does it require close interaction with organisms, but it also requires close collaboration with different people (SANTOS, 2003).

Several research institutions are setting up biological collection programs as part of their scientific strategic plan. Some of these research institutions are: the Global Biodiversity Information Facility (GBIF) (GBIF.ORG,), the Biodiversity Database Collection of the National Research Institute for the Amazon (INPA) (INPA.BR,), the Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) (LBA.BR,), the Reference Center on Environmental Information (CRIA) (MCTIC, 2017), and the New York Botanical Garden (NYBG) (NYGB, 2017).

While most researches in biodiversity pay much attention to the generation of biodiversity datasets and Web access, information that specifies how/where these data are derived and who owns/publish the data is often ignored.

In this paper, we propose a conceptual model for provenance in biodiversity data for species identification. This model is based on W3C PROV Data Model (MOREAU; Missier (Eds.); W3C Provenance Working Group, 2013a). The PROV specification provides the concepts and supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web (TAXIDOU *et al.*, 2015). This conforms to the

Figure 9 – Example of provenance in biodiversity data.

Source: Elaborated by the author.

principles of Linked Open Data (LOD), which encourages a set of best practices for publishing and connecting structured data on the Web (KUHN; KAUPPINEN; JANOWICZ, 2014).

To further clarify the relation between the biodiversity domain and provenance, we provide an example in Figure 9.

Due to new discoveries, species names could change over time. As Figure 9 shows, this is the case for the *Batrachospermum Alga*. Data related to this species was collected and saved into a csv file by the Collector. After a cataloguing process, this species was determined as *Batrachospermum Helminthosum* by the Cataloguer. After 15 years, a User needs to determine the genetic name of this species. Subsequent to molecular studies, this species had its name changed to *Batrachospermum viride-brasiliense*. This is a common problem faced by biodiversity collections. The user needs to answer the following questions: Who was the cataloguer of the species? Who was the collector of the species? When was the data collected? Why was the data collected? Which institution can provide the data? The user needs to know the history (provenance) of the species. This means that the trustworthiness of the cataloguer, the person who determined the species and the user involved should be judged, since they participate in the identification and modification of the species name.

As our main contribution in this paper, we present an extended provenance model applied to biodiversity datasets. We mapped a set of representative data about biodiversity (217,829 records) from the Botanical Institute (IBt/SP). This data was downloaded from the SpeciesLink web site¹. SpeciesLink is a distributed information system that integrates primary data from biological collections. We also use the GeoSPARQL language² (an extension to the SPARQL

¹ <http://splink.cria.org.br/>

² <http://www.opengeospatial.org/standards/geosparql>

language that allows queries based on spatial relations, such as being inside a polygon, etc.) to answer spatially complex queries.

The remainder of this article proceeds as follows: Section 2.3 discusses related work. Section 2.4 shows the provenance model for biodiversity datasets. In Section 2.5, we present a use case, where we model all 217,829 records of the SpeciesLink website using our proposed approach and Section 2.6 concludes by summarizing our results and describing future work.

3.3 Related work

Biodiversity data is an assortment of different types of data about organisms that co-occur in time and space (geospatial) (MAGNUSSON, 2013). We investigated existing works related with the geospatial and biodiversity domains that use a provenance model.

Zhao et al. (ZHAO *et al.*, 2010) propose a method for recording the provenance data into biological datasets. This method helps scientists obtain the information about particular biological terms. The authors use the Dublin Core Metadata Initiative (DCMI) (WEIBEL; KOCH, 2000) and named Resource Description Framework (RDF) graphs to represent the aspects of data provenance. The authors only considered two types of links between a pair of genes, i.e., either they are the same or different from each other. Provenance information about these links is needed to provide reliable and accurate services to researchers.

Sousa et al. (SOUSA *et al.*, 2014) propose a provenance-based approach to manage long term preservation of scientific data. This approach uses a case study related to the long term preservation of the animal sound collection at the Fonoteca Neotropical Jacques Viellard (FNJV)³. Their approach is based on the Open Provenance Model (OPM) (MOREAU *et al.*, 2008). However, this approach does not provide support to connect curated metadata with Linked Open Data. It would allow breaking down disciplinary boundaries among repositories and enhance reuse.

There are a number of studies, that have used provenance in the geospatial domain (WANG *et al.*, 2008; YUAN *et al.*, 2013; MALAVERRI; MEDEIROS; LAMPARELLI, 2012; CELINO, 2013). For example, Wang et al. (WANG *et al.*, 2008) propose a provenance-aware architecture to record the lineage of spatial data in Geographic Information Systems (GIS). Their architecture is based on the OPM model and organize spatial provenance as objects in a spatial data store, while a spatial process provenance is represented as graphs and stored using Semantic Web technologies based on the RDF, and is backed with standard storage technologies (e.g. database) and RDF stores (e.g. Sesame). However, the authors do not consider the variable spacetime to enhance the geospatial capabilities of either provenance-aware GIS.

Yuan et al. (YUAN *et al.*, 2013) propose a Linked Data approach for geospatial data

³ <http://www2.ib.unicamp.br/fnjv/>

provenance. The authors defined a geospatial data provenance ontology based on the Provenir ontology (SAHOO, 2009), published geospatial data provenance as Linked Data, and analyzed queries of linked geospatial data provenance. Their approach is based on the Registry Information Model (ebRIM)⁴ and the DCMI model. However, this approach does not achieve geospatial reasoning based on the linked geospatial data provenance.

Magnuson et al. (MAGNUSSON, 2013) explain that biodiversity research will often have specific taxonomic or ecosystem interests, but the primary keys that link all of these things are related to space (geographic location) and time (when the observation was made). However, most of the available biological collections are not consistently georeferenced making use of a coordinate system. The authors explain that there is still a fundamental lack to answer complex queries, i.e., queries that need logical inference that use spatial and temporal relations (e.g., plantations within a protected area in Manaus, Brazil between 2005 and 2011).

A critical look at the available literature indicates that a number of techniques have been developed for using provenance models, such as OPM and DCMI, in different scientific domains (biological, biodiversity and geospatial). Despite the variety of models, there is currently no unified, conceptual model for biodiversity information that can be applied to different datasets and setups, while remaining both expressive and generic enough to cover many use cases.

The PROV specification (MOREAU; Missier (Eds.); W3C Provenance Working Group, 2013b) defines a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or thing in the world. However, there is a lack of expressiveness using this generic W3C recommendation to model the different types of organisms that co-occur in time and space (geospatial relations).

3.4 Architecture for publishing linked biodiversity data

This section presents our architecture for publishing linked biodiversity data (as illustrated in Figure 10). Our architecture uses Linked Data and Semantic Web standards (Resource Description Framework RDF (LASSILA *et al.*, 1998) and the Web Ontology Language OWL (ANDRADE; BAPTISTA; DAVIS CLODOVEUAUGUSTO, 2014)) to represent biodiversity data.

When biodiversity data are collected and catalogued by third parties (*Cataloguer and Collector*), they are registered and stored in commercial spreadsheets (e.g., Microsoft Excel) or databases (e.g., DBase, PostgreSQL) or files associated with statistical programs (e.g., R or SPSS) (MAGNUSSON, 2013). A mapping component converts the biodiversity data to RDF. The RML language (DIMOU *et al.*, 2014) is used to map biodiversity data to RDF triples. The RDF triples are integrated with our provenance model for biodiversity data. These RDF triples are stored in triple stores. After that, users can retrieve biodiversity information through SPARQL

⁴ <http://docs.oasis-open.org/regist/regist-core/v4.0/regist-core-rim-v4.0.html>

Figure 10 – Architecture of publishing linked biodiversity data.

Source: Elaborated by the author.

queries. In the next subsections, we explain on more details each component of our architecture for biodiversity data provenance.

3.4.1 Provenance model for biodiversity data (BioProv)

In order to create our provenance model, five biodiversity scientists were interviewed to categorize important information from biodiversity data (e.g. collecting process, genus, family, species, location). These interviews helped us to understand more about their work and to form a common ground for discussions. A list of our interviews are available at https://www.researchgate.net/publication/301287330_Interviews.

Using the information gathered through these interviews, we created our provenance model for biodiversity data, as shown in Figure 11. This model is based on the W3C PROV Data Model (MOREAU; Missier (Eds.); W3C Provenance Working Group, 2013b). It defines a set of starting point terms divided into three classes: *Entity*, *Agent* and *Activity*. These classes are associated by relations such as *prov:wasAttributedTo*, *prov:wasInformedBy*, etc. The entity responsible for commanding the execution of an activity is modeled as an *Agent*.

In our model, a species is denominated *Collection Object (CO)*. The CO was generated by an activity denominated *Collecting* (in Figure 11, *prov:wasGeneratedBy*). A *Collector* agent was associated with this activity (in Figure 11, *prov:wasAssociatedWith*). We trace the date this activity was executed with the property *prov:atTime*.

After a *Collecting* process, the *Collection Object* needs to be identified through an activity denominated *Cataloguing* (in Figure 11, *prov:wasGeneratedBy*). The *Cataloguer* agent assigns a unique identifier to each collection item using the taxonomic classification. The *Cataloguer* uses the reference work to indicate the published material in which the collection object is mentioned (in Figure 11, *Reference Work*).

The *Collection Object* has a relationship to the *Reference Work*, indicating the published

Figure 11 – Provenance model for biodiversity data.

Source: Elaborated by the author.

material in which the *Collection Object* is mentioned (in [Figure 11](#), *prov:wasDerivedFrom*). The *Location* describes where the data was collected (in [Figure 11](#), *prov:atLocation*). It defines locality, named place, habitat and spatial information.

The *Agents* describe persons and organizations, which deal with the biological collection information and interacts with all activities and all information that is updated in the model. The *Curator*, *Collector*, *Cataloguer* and *User* agents are members of an *Organization Agent* (in [Figure 11](#), *prov:actedOnBehalfOf*). An *User* agent can be influenced by a *Cataloguer* agent and vice-versa (in [Figure 11](#), *prov:wasInfluencedBy*), since they participate in the identification and modification of the species name.

The PROV data model provides extensibility points that allow designers to specialize it for specific applications or domains (subtypes, roles, and Attribute-value lists) (MOREAU; Missier (Eds.); W3C Provenance Working Group, 2013a). In order to model the species names that are defined by the agents, we propose the following extensions that are subtypes of *prov:Entity*:

- *bioprov:OriginalSpeciesName*: denotes an original species name that is not derived from any other name and the collector who emitted it is the initiator of the identification process.
- *bioprov:CataloguerSpeciesName*: denotes a species name which is based on another name that has been published in the past. The *Cataloguer* agent emitted this name based on the original species name.
- *bioprov:MolecularSpeciesName*: denotes a species name that is produced by modifying an existing species name. It is possible that the species name is altered.

With these three extensions we have covered the main case of the example illustrated in Section 1 (species names could change over time). Our model is available at https://www.researchgate.net/publication/299690682_BioProv.

3.4.2 Mapping provenance and biodiversity data to RDF

The Mapping component of our architecture to publishing linked biodiversity data loads the domain ontologies, provenance model, taxonomic information and the collection database and transforms them in a set of RDF triples. We used the RDF Mapping Language (RML) (DIMOU *et al.*, 2014) to represent the mapping between rows of data tables (in csv files) and properties and objects in RDF.

RML is a mapping language defined to express customized mapping rules from heterogeneous data structures and serializations to the RDF data model. RML is defined as a superset of the W3C-standardized mapping language (R2RML) (DIMOU *et al.*, 2014). A Triples Map defines how triples of the form (subject, predicate, object) will be generated. A Triples Map con-

sists of three main parts: the Logical Source, the Subject Map and zero or more Predicate-Object Maps. In the following Source Code 2, we show an example of a triple map:

Source code 2 – Example of a triple map.

```

1: @prefix rr:<http://www.w3.org/ns/r2rml#>.
2: @prefix rml:<http://semweb.mmlab.be/ns/rml#> .
3: @prefix geobio:<http://geobio.lod.usp.br/>.
4: @prefix prov:<http://www.w3.org/ns/prov#> .
5: @prefix dcterms:<http://purl.org/dc/terms/> .
6: @prefix adms:<http://www.w3.org/ns/adms#>.
7: @prefix bioprov:<http://geobio.lod.usp/bioprov/>.
8: @prefix skos:<http://www.w3.org/2004/skos/core#> .
9: @prefix xsd:<http://www.w3.org/2001/XMLSchema#> .
10: @prefix foaf:< http://xmlns.com/foaf/0.1/> .
11: <#BioMapping>
12: rml:logicalSource[rml:source "speciesLink.csv";
13: rml:referenceFormulation ql:CSV];
14: rr:subjectMap[
15: rr:template "http://geobio.lod.usp.br/ibt/id/{code}";
16: rr:class adms:Identifier];
17: rr:predicateObjectMap [rr:predicate skos:notation;
18: rr:objectMap[ rml:reference "code";]];
19: rr:predicateObjectMap[ rr:predicate prov:agent;
20: rr:objectMap[ rml:reference "institutioncode";]];

```

The Logical Source represents the source to be mapped. This can be a pointer to any dataset (Line 12-13). The Subject Map (Line 14-16) defines how unique identifiers (URIs) are generated for the mapped resources and is used as the subject of all RDF triples generated from this Triples Map. A Predicate-Object Map (Line 17-20) consists of Predicate Maps, which define the rule that generates the triple's predicate and Object Maps or Referencing Object Maps (Line 18, 20), which define how the triple's object is generated. The Subject Map, the Predicate Map and the Object Map are Term Maps, namely rules that generate an RDF term (an Internationalized Resource Identifier (IRI), a blank node or a literal).

3.5 Use case

In order to validate our provenance model, biodiversity scientists were interviewed to define use cases with features and scenarios to identify the various user tasks. A list of our use cases are available at https://www.researchgate.net/publication/301287330_Interviews. In this article, we present one of these use cases:

- USE CASE: Molecular Identification of *Cladophora delicatula* Alga
- USER: Monica Paiano, 32 years-old, Collector and Cataloguer of the Laboratory BETA, UNESP, Brazil and Phycology Research Group, Ghent University, Belgium.
- GOAL: To determine the scientific name of *Cladophora delicatula* alga through a genetic identification.
- MOTIVATION: Due to new discoveries, species names of *Cladophora delicatula* alga could change over time. Keeping such data up to date and consistent is extremely important because the presence or not of some species of this alga can serve as biological markers (bio indicators) that indicate the degree of conservation or degradation in a aquatic habitat.
- TASKS
 1. Retrieve all information about *Cladophora delicatula* alga. For example, when it was collected, who collected it, all the specific characteristics;
 2. Store all the information in csv, text file or in a biodiversity database;
 3. Start the molecular studies of the *Cladophora delicatula* alga;
 4. Identify the species names in a flexible way: using the broader taxonomic level (*phylum or genus*) without having to worry about whether the original collection used this particular classification level.
- NECESSARY TOOL FEATURES
 1. Retrieve all specifications of the bio-marker species using the species name or any higher taxonomic level, like phylum, genus or family.

After studying our use case, we mapped the corresponding biodiversity provenance records to RDF. We used the RML language to convert all IBT's records from the SpeciesLink web site (217,829 records) to RDF triples. This RDF data was stored in our Strabon Triple Store and can be explored using SPARQL queries. The biodiversity datasets are available at https://www.researchgate.net/publication/299740010_ProvGeoIBT_Dataset.

In order to show how the previous use case was implemented, in the following subsections we explain the more important activities of our provenance model: *Collecting and Cataloguing Activity*.

3.5.1 Collecting activity

For this activity ([Figure 11](#), *Collecting*), it is important to keep track of (1) When was the species collected (2) Who was the collector of the species, (3) Where was the species collected, and (4) Which institution can provide the species. This activity is crucial for capturing the origin

of the biodiversity data, as it is only at this step that information is known. In the following, we show an example of our provenance model applied to the collecting process for *Cladophora Delicatula* species.

In Source Code 3, we used the subclass *bioprov:OriginalSpeciesName* to define the species name ([Figure 11](#), *bioprov:OriginalSpeciesName*).

Source code 3 – Example of *Collecting* Activity.

```

1: :AgentCollector
2: a prov:Agent;
3: foaf:givenName "M.C.Marino & R.Marino";
4: prov:actedOnBehalfOf :IBT.
5:
6: :Collecting
7: a prov:Activity;
8: prov:wasAssociatedWith :AgentCollector;
9: prov:atTime "1966-11-11T01:01:01Z";
10: bioprov:OriginalSpeciesName "Cladophora Delicatula".

```

In Source Code 4, we used the properties *prov:activity*, *prov:atTime* and *prov:atLocation* to define the spatiotemporal location of our species collected. To deal with this, we used the GeoSPARQL language and the Well-Known Text (WKT), a pattern defined by the Open Geospatial Consortium (OGC) for defining coordinates in the form of points, lines and polygons.

Source code 4 – Example of spatiotemporal location.

```

1: :Location
2: a prov:Entity;
3: foaf:name "Ponta do Gil Lake";
4: prov:qualifiedGeneration [ a prov:Generation;
5: prov:activity geo:feature;
6: prov:atTime "1966-11-11T01:01:01Z";
7: prov:atLocation
8: "POINT(-46.7175 -23.653056)"^^geo:wktLiteral;] .

```

3.5.2 Cataloguing activity

The Cataloguing activity ([Figure 11](#), *Cataloguing*) permits the taxonomic identification of the biodiversity data. The taxonomic identification information contains the identifiers of the biological classification as Order, Family, Genus, Species and nearly all of them were used. In the following Source Code 5, we show an example:

Source code 5 – Example of *Cataloguing* Activity.

```

1: :AgentCataloguer
2: a prov:Agent;
3: foaf:givenName "D.P. Santos";
4: prov:actedOnBehalfOf :IBT .
5:
6: :Cataloguing
7: a prov:Activity;
8: prov:wasAssociatedWith :AgentCataloguer;
9: prov:atTime "1982-01-01T01:01:01Z";
10: bioprov:CataloguerSpeciesName "Cladophora Delicatula"

```

We use the ProvValidator and ProvTranslator tools⁵ to validate and translate PROV representations about the collecting and cataloguing activities of our biodiversity datasets, making them fully interoperable. The complete representations of our use case are available at https://www.researchgate.net/publication/301287278_BioProvExample.

To integrate the biodiversity data in RDF to the wider LOD community on the Web, we set up a SPARQL endpoint⁶. Our endpoint allows third-party programs to query our knowledge base, via the SPARQL language, and reuse it in their applications.

3.5.3 Querying linked biodiversity data provenance

For the previous use case, a User wants to identify all information about *Cladophora delicatula* alga. One of the big advantages of having the biodiversity data in RDF is to be able to connect it to other sources. We created a SPARQL query for integrating different triples stores. In the following Source Code 6, we show the SPARQL query used to obtain the provenance information of a specific dataset.

Source code 6 – SPARQL query used to obtain the provenance information of a specific dataset.

```

1: SELECT ?Species ?agent ?activity ?date ?pontowkt
2: WHERE {?Species prov:wasAttributedTo ?agent.
3: ?Species prov:wasGeneratedBy ?activity.
4: ?activity prov:atTime ?date.
5: ?Species geo:hasGeometry ?Geometry.
6: ?Geometry geo:asWKT ?pontowkt.
7: ?Species bioprov:CataloguerSpeciesName
8: "Cladophora delicatula".
9: FILTER(?date > "1980-01-01T01:01:01Z").

```

⁵ <https://provenance.ecs.soton.ac.uk/>

⁶ <http://java.icmc.usp.br:2190/>

```

10: geof:sfWithin(?pontowkt ,
11:   "Polygon(-1 -58,-7 -58,-7 -69,-1 -69,-1 -5))").}

```

Using this query, we could retrieve the lineage of the *Cladophora delicatula* species. In our provenance model, we reused the GeoSPARQL ontology terms to describe georeferenced data. This implementation permits to answer complex queries such as: Locate all occurrences containing *Cladophora delicatula* alga samples inside of a Polygon (-1 -58, -7 -58, -7 -69, -1 -69, -1 -58)

3.6 Conclusion

In this work, we presented a model for biodiversity data provenance (BioProv). This model is based on the W3C PROV ontology and data model. BioProv enables applications that analyze biodiversity to incorporate provenance data in their information. We defined a mapping document for the biodiversity data from IBT to generate RDF triples. We also reused the GeoSPARQL ontology terms to describe georeferenced data. We use the provenance information to allow experts in biodiversity to perform queries and answer scientific questions.

As future work, we also intend to extend our current implementation with more advanced structured queries, in partnership with biodiversity researchers. We also intend to build a benchmark to evaluate the precision and recall of our queries.

3.7 Remarks about the chapter

In this chapter, we responded the research question 3, "How can provenance be modeled for multiple biodiversity use cases?", by proposing a model for biodiversity data provenance (BioProv). This model is based on the W3C PROV ontology and data model. The PROV data model provides extensibility points that allow designers to specialize it for specific applications or domains. Thanks to its generic core specification, multiple use cases (e.g. for the biodiversity domain) related to provenance can be modeled in an interoperable way. This means that researchers no longer have an excuse to lock in their provenance in a way that only they can understand.

In order to model the species names that are defined by the Cataloguers, we proposed three extensions to the PROV ontology subclasses of the `prov:Entity` class: *bioprov:OriginalSpeciesName*, *bioprov:CataloguerSpeciesName* and *bioprov:MolecularSpeciesName*. We used the ProvValidator and ProvTranslator tools⁷ to validate and translate PROV representations about the collecting and cataloguing activities of our biodiversity datasets, making them fully interoperable.

⁷ <https://provenance.ecs.soton.ac.uk/validator/view/translator.html>

We presented a use case, where we model all 217,829 plant records, of the SpeciesLink web site, using our approach.

We demonstrated that using BioProv, we can model provenance in the biodiversity domain for the use case of Molecular Identification of *Cladophora delicatula* Alga. We also showed that BioProv can model different types of organisms that co-occur in time and space (geospatial relations).

In the next chapter, we will continue to improve our architecture adding, besides the BioProv model, spatiotemporal features.

LINKING BIODIVERSITY DATA USING SPATIOTEMPORAL INFORMATION

In this chapter, the research question 1, "How can we integrate heterogeneous biodiversity semantic data using their spatiotemporal information and provenance?", is answered. The chapter's article, entitled "STBioData: Linked Biodiversity Data using their Spatiotemporal Information, a Use Case about the Conservation Status of Amazon River Dolphins", proposes the STBioData architecture and prototype system for linked biodiversity data using the BioProv model (described in chapter 3). It uses spatiotemporal information to integrate heterogeneous biodiversity data. The chapter is organized as follows. Section 4.1 describes the motivation behind the STBioData architecture. Section 4.2 describes the related work. The architecture for linked biodiversity data is presented in Section 4.3. Section 4.4 presents our experiments and use case. Section 4.5, concludes by summarizing our results and describing future work. Finally, the remarks of this chapter are given in Section 4.6.

4.1 Introduction

Linked biodiversity data has recently received much attention, as researchers have started tapping the wealth of biodiversity information available on the Web. As a result, in the last few years, the Web of data is being rapidly populated with biodiversity information. The Global Biodiversity Information Facility(GBIF) (GBIF.ORG,) is the first institution that has made various kinds of biodiversity data available as open Linked Data. GBIF has proposed to the W3C the creation of the Vocabulary Management Task Group (VoMag) (COMMUNITY, 2017). Its task is to develop best practices and guidelines for maintaining Resource Description Framework (RDF) ([LASSILA *et al.*, 1998](http://LASSILA)) vocabularies of terms and concepts for biodiversity informatics. In March 2017, there were 733,827,815 data records (352,593,699 georeferenced) accessible from the GBIF portal.

Other similar effort is the Linked Brazilian Amazon Rainforest Dataset (LBARD) ([KAUPPINEN et al., 2014](#)). It divides the Brazilian Amazon forest in squares of 25 by 25 Km and, for each square, provides data about deforestation, rivers, road network, population and economic activity (including cattle numbers and market prices of agricultural products). LBARD represents its data as RDF triples accessible over the web. Efforts, such as GBIF and LBARD, only provide data. If users need to visualize Amazon Rainforest data they have to use programmatic solutions.

Given its nature, biodiversity data usually provide detailed information about specimen species, their distribution in space (where items were located) and time (when something was collected). This spatiotemporal information can be combined with other geo-referenced data for the creation of predictive distribution maps ([SANTOS, 2003](#)) .

Maps are a very convenient way to visualize biodiversity information. To generate these maps, biodiversity experts have to recover the relevant information from different biodiversity repositories, organize it and plot the results as maps. To do that, given the large quantity of datasets generated by research institutions, they use programmatic tools, such as the R language ([R Development Core Team, 2008](#)) from the R Project (software environment for statistical computing and graphics). R is becoming a popular language for working with biodiversity data, but this work is manual and time consuming.

To further clarify this problem, we use, as an example, the process of identifying the conservation status of species over time. The conservation status of a species is an indicator of how likely it is that it will survive at the present or in the near future. [Figure 12](#) shows this process for the case of a dolphin from the *Inia Geoffrensis* species. It is a process that happens over time.

This dolphin species is at risk of extinction. It is important for biodiversity specialists to know its conservation status. In the process, shown in [Figure 12](#), data related to occurrences (collections or observations) of specimens of this species are collected and saved in a csv file by *Collectors*. After a *cataloguing* process, the data are confirmed to belong to specimens of the *Inia Geoffrensis* species by *Cataloguers*. In some future point in time, a researcher needs to determine the conservation status of this species over a period of time. To do that, the researcher needs to compare the known geographic distributions of *Inia Geoffrensis* dolphins, over the years, to occurrence data for its specimens. In this way, he can confirm (or not) the species distribution, and thus its conservation status, over the years.

As [Figure 12](#) shows, this process can span over long time periods and involve many people. To ensure data integrity and trust, provenance data is added. Data provenance, also called data lineage, records sources as well as a set of processing steps applied to sources ([CELINO, 2013](#)) . Using this information, users can access data history (provenance) and determine, for instance, if they trust the *cataloguer*, the person who determined the specimen's species ([AMANQUI et al., 2016b](#)).

Figure 12 – Example of the process to identify the conservation status of species and plot it in a map.

Source: Elaborated by the author.

Conservation status data and geographic location of *Inia Geoffrensis* dolphins can be found in the IUCN Red List of Threatened Species (IUCN) ([SPECIES, 2017](#)). These data show the geographic areas where the species could be found (its range) in each year (using polygons in ESRI shapefile format). After that, the researcher has to invest a considerable amount of time, programming in R (or other computer language) and performing several manual tasks: to integrate the spatiotemporal occurrence information, from the csv files and the ESRI shapefiles, to generate maps showing both information over time. If he does not know R (or another language), he is out of luck.

This is a common problem in the biodiversity domain. How to integrate spatiotemporal information from heterogeneous sources? How to know the origin (provenance information) of data about species? How to download this information to integrate with information from other domains?.

As our main contribution in this chapter, we present a novel architecture to automatically link spatiotemporal biodiversity data, from heterogeneous data sources, to enable easy searching, visualization and downloading of relevant data. It supports the generation of interactive maps and mapping between biodiversity data and the ontologies describing them.

A web based prototype based on this architecture was implemented, the STBioData system. It allows users to automate the conservation status identification task, shown in the right of [Figure 12](#). It uses collection data, from GBIF, and species geographic distributions and conservation status, from IUCN. These data are converted to Linked Data, enriched and saved

as RDF Triples. Users can access the system, using a web interface, and search for collection and distribution records based on species names and time ranges. After a data set is recovered, it can be displayed in a geographic visualization. The records contents are also shown (including provenance data) together with links to the original records at GBIF and IUCN. When a user is happy with the data he got, he can export it as a CSV or RDF file or get a print out in PDF. Choosing different time periods, users can verify the evolution of a species distribution.

It was tested using a representative biodiversity dataset: 46568 records from the National Institute for Amazonian Research (INPA) ([INPA.BR](#),), the Brazilian Agricultural Research Corporation (EMBRAPA) ([EMBRAPA.BR](#),) and the IUCN Red List of Threatened Species (IUCN) ([SPECIES](#), 2017). This dataset was downloaded from the SpeciesLink ([INPA](#), 2017) and IUCN web sites. SpeciesLink is a popular repository of Brazilian biodiversity data.

This prototype was evaluated by users from the Laboratory of Molecular Biodiversity and Conservation of the Federal University of São Carlos (UFSCar) and Semantic Web specialists from the University of São Paulo. We conducted an exploratory evaluation, using questionnaires, to find out if our architecture helped biodiversity specialists retrieve and link biodiversity information. 90% of the specialist did not have problems understanding the system's interface and all of them were able to describe the purpose of the STBioData system.

We also compared the task of finding relevant species data for the *Inia geoffrensis* dolphin in the SpeciesLink site and in the STBioData system. Biologists from UFSCar conducted this experiment. Our results show that the STBioData tool had better precision (22% improvement) and recall (24% improvement) when compared to the keyword based search, used by SpeciesLink.

We also show that, because the biodiversity data is now in a Semantic Web format and mapped to ontology terms, it is easier to link it with information from other sources.

Using the STBioData, we were able to determine the conservation status of Amazon river dolphins (*Inia geoffrensis*). Using it, biodiversity specialist will be able to investigate the conservation status of Amazon river dolphins to support the process of creating protected natural areas.

The remainder of this article proceeds as follows: Section 4.2 discusses related work. Section 4.3 shows the architecture for linked biodiversity data. Section 4.4 presents our experiments and use case. Finally, Section 4.5 concludes by summarizing our results and describing future works.

4.2 Related work

We conducted a systematic review in the literature for processes that could automate conservation status identification. We could not find any that could deal with time based search, RDF conversion, provenance and visualization for biodiversity data. So, we decided to focus on

the three main techniques used by STBioData: Linked Data in the biodiversity domain, geospatial visualization and provenance of linked biodiversity data. We found some works, but none used the three techniques together. Below we list them:

The article of Cardoso et al. (CARDOSO *et al.*, 2015) is the most related to our work. It describes a geographic gazetteer that associates place names to geographic coordinate data. It used data from two large biodiversity repositories: GBIF and SpeciesLink. Skevakis et al. (SKEVAKIS *et al.*, 2014) proposed an architecture to link biodiversity data from the Natural Europe History Museums with Geonames to find geographical information. Moura et al. (MOURA *et al.*, 2015) proposed an architecture to link information from the Guanabara Bay ecosystem to the Darwin Core ontology. However, these efforts do not address the temporal dimension and cannot answer complex queries with both spatial and temporal characteristics, for example “find farms within a certain amount of protected area between 2008 and 2016”.

We investigated existing works that allow the geospatial visualization of Linked Data. Map4RDF (LEON *et al.*, 2012) is a mapping and faceted browsing tool for exploring, visualizing and interacting with RDF datasets enhanced with geospatial information. This tool provides geospatial and geometrical visualization using Google Maps, Open Street Maps and national spatial data infrastructures. Map4RDF allows the visualization of GeoSPARQL and WKT geometries such as polygons, points, etc. The Simple Protocol and RDF Query Language (SPARQL) is a W3C standard language for querying, retrieving and manipulating RDF data (triples) (PRUD'HOMMEAUX; SEABORNE, 2006). GeoSPARQL is a SPARQL extension that allows queries based on spatial relations and the representation of coordinates in the form of points, lines and polygons. Well-known text (WKT) is a text markup language for representing vector geometry objects on a map, spatial reference systems and transformations between reference systems. Facete (STADLER; MARTIN; AUER, 2014) is a Web application for the exploration of RDF geospatial data. It provides domain independent faceted filtering capabilities operating directly on SPARQL. SexTant (NIKOLAOU *et al.*, 2015) is a web-based tool to visualize spatiotemporal information on a map. This tool is based on the Open Geospatial Consortium (OGC) standards: the OpenGIS Simple Features Access (OGC-SFA) standard (INC., 2011), the Geography Markup Language (GML), the KML file format and the GeoSPARQL query language (CONSORTIUM, 2017). However these tools do not help when a user wants to search, retrieve and visualize temporal information and relate information to other datasets. For example, in the case of a dolphin from the *Inia Geoffrensis* species: “find this dolphin within an Atlantic ocean protected area between 2008 and 2016”.

Provenance in Linked Biodiversity Data has been a well researched topic. A review of relevant models focused on provenance assessment of LOD can be found in (ZAVERI *et al.*, 2016). Until now, the focus of provenance research has been on the design of models with specific goals, for example: Hasnain et al. (HASNAIN *et al.*, 2017) proposed BioFed, a user friendly system for federated SPARQL query processing based on biological data. It provided basic

provenance information: the sources queried, the number of triples returned and the retrieval time. Gaignard et al. (GAIGNARD *et al.*, 2014) proposed an integrated provenance approach in which Life-Science knowledge is captured through domain ontologies and linked to Life-Science data analysis tools. Until now, the focus in previous work has been to extend the DCMI and OPM provenance models with specific goals (e.g. linked geospatial data provenance, record provenance data into biological datasets). However these models do not allow extension of their classes to convey time information. In (AMANQUI *et al.*, 2016b), we proposed the BioProv Model, which is a new biodiversity provenance model that extends the W3C PROV Data Model with temporal information.

Despite the variety of biodiversity works found in our bibliographic research, we did not find a unified conceptual model to represent biodiversity information that can be applied to different datasets and setups, while remaining both expressive and generic enough to cover many use cases. Most of the works do not address the temporal dimension and cannot answer complex queries with both spatial and temporal characteristics. Also, these works do not offer solutions to search, retrieve and visualize geographical information of species. There is no tools to links biodiversity datasets to other domains that consider a provenance model with spatiotemporal information. Also, there are no approaches to visualize linked biodiversity data that have spatiotemporal relations. It is important to remember that biodiversity information is also stored in geometries, such as points and polygons, and that it is necessary to also link these geometries when enriching data.

To address these problems, this chapter proposes a novel Linked Biodiversity Data Architecture. It uses ontologies to integrate biodiversity datasets, with datasets from other domains, to improve retrieval, linking, and visualization of biodiversity information. It allows the mapping of biodiversity data, from heterogeneous sources, to terms from relevant ontologies, such as Darwin Core (BASKAUF; WEBB, 2016), DBpedia (LEHMANN *et al.*, 2015) and GeoSPARQL (CONSORTIUM, 2017). This information is stored using Semantic Web formats and queried using Semantic Web tools (such as triple stores). This architecture supports the generation of interactive maps and also allow users to download information in two formats : CSV and RDF.

Table 1 summarizes the differences between our approach and the ones found in the literature. It highlights which works use ontologies, provenance, spatiotemporal queries, LOD, visualization of species occurrences in a map, semantic search, online statistics and RDF download. A critical look at the available literature indicates that, while most researches in the biodiversity domain pay much attention to the generation of biodiversity data in RDF format, information that specifies how/where/when these data were derived is often ignored (AMANQUI *et al.*, 2016a). We also still lack:

- Links from biodiversity datasets to others domains.

- Visualization of linked biodiversity data using spatial and temporal relations.
- Links from biodiversity information, stored in geometries, to other data (to enrich them).

Table 1 – Summary of the features present in related works. OK indicates feature support, partial indicates partial support, and ‘-’ indicates no support. UO: Use Ontologies, UP: Use Provenance, STQ: SpatioTemporal Queries, OVM :Occurrence Visualization in a Map, SO: Search Occurrences, OS: Online Statistics, LOD: Linked Open Data.

Related Work	UO	UP	STQ	LOD	OVM	SO	OS	Download RDF
Cavoto et al.	OK	-	-	OK	-	OK	-	-
Hasnain et al.	OK	OK	-	OK	-	partial	-	-
Stucky et al.	OK	-	-	OK	-	partial	-	partial
Fafalios et al.	OK	-	-	OK	-	OK	-	OK
Moura et al.	OK	-	-	OK	-	-	-	partial
Skevakis et al.	OK	-	partial	OK	-	partial	-	partial
SWI Gazeeteer	partial	-	partial	OK	OK	partial	-	partial
BioProv	OK	OK	partial	OK	-	partial	-	partial
SpeciesLink	-	-	-	-	OK	OK	OK	-
GBIF Portal	OK	-	partial	partial	OK	OK	OK	partial
STBioData	OK	OK	OK	OK	OK	OK	OK	OK

Source: Elaborated by the author.

In the next section we introduce our Linked Biodiversity Data Architecture that tackles the problems mentioned above.

4.3 The STBioData architecture

This section presents the Linked Biodiversity Data Architecture for the STBioData prototype. [Figure 13](#) presents the architecture’s overall schema. It is divided in layers:

- The Presentation layer is responsible for the interaction between users and STBioData system. A user interaction begins with an initial search using a keyword list and date range, entered by the biodiversity specialist, which represents his/her search intentions.
- The Business layer processes the information retrieval through the query reformulation component. Querying the semantic data is simplified because of the use of triples. These triples are generated when database records are mapped into ontology concepts through the mapping mechanism.
- The Data layer provides access to the RDF triples stored in the Triple Store (in our prototype, GraphDB) to the layers above it and to software agents in the Web (through a SPARQL Endpoint).

- The Model layer contains a collection of concepts and relationships that characterize the knowledge in the biodiversity domain (domain ontologies).

Figure 13 – Linked biodiversity data architecture.

Source: Elaborated by the author.

The STBioData architecture layers use four main components:

1. Domain Ontologies: they contain a collection of concepts and relationships that characterize the biodiversity domain knowledge;
2. The Mapping Component: it maps collection tabular records to RDF format;
3. The Web Interface, which process queries from users and visualize the information result in a interactive map;
4. The Query Reformulator: it plays a central role in the architecture by processing input query data.

4.3.1 Domain ontologies for biodiversity

Ontologies contain a collection of concepts and relationships that characterize the knowledge in a domain (AMANQUI *et al.*, 2014). They provide a common reference point that support the semantic integration and interoperability of biodiversity data. We reused the following ontologies:

- The Darwin Core Standard (BASKAUF; WEBB, 2016) is a biological vocabulary that includes a glossary of terms intended to facilitate the sharing of information about biological diversity, by providing reference definitions, examples, and commentaries. In our architecture, we use Darwin Core to describe the taxonomy information of the species (e.g. kingdom, class, genus, etc), properties, elements, fields, columns, attributes and concepts, for example: *dwc:institutionCode*, *dwc:collectionCode*, *dwc:catalogNumber*, *dwc:scientificName*, *dwc:basisOfRecord*, etc.
- The GeoSPARQL Ontology (CONSORTIUM, 2017) is an ontology that allows queries based on spatial relations, such as being inside or to border, and the representation of coordinates in the form of points, lines and polygons. Our architecture uses this ontology to describe the spatial information of each species occurrence. For example: *geo:hasGeometry*, *sf:Point* and we use the WKT format to represent point coordinates, for example: (Point(lat,long), *geo:wktLiteral*). We also use GeoSPARQL to answer more complex queries, such as Locate all occurrences containing *Inia geoffrensis* records inside a polygon in the Amazon Forest.
- The DBpedia Ontology (LEHMANN *et al.*, 2015), is a community-curated ontology consisting of 320 classes that form a subsumption hierarchy and are detailed by 1,650 different properties. The ontology is maintained and extended by the community in the DBpedia Mappings Wiki. This community also creates mappings from Wikipedia information representation structures (info-boxes created by Wikipedia editors) to the DBpedia ontology. Our architecture uses instances from the English (dbpedia.org) and Portuguese (pt.dbpedia.org) DBpedias. The Portuguese version is mainly used to describe Brazilian geography.
- The BioProv Model (AMANQUI *et al.*, 2016b) is a conceptual model for provenance in biodiversity data based on the W3C PROV Data Model (MOREAU; Missier (Eds.); W3C Provenance Working Group, 2013b). The PROV model provides the concepts and supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments, such as the Web. Our architecture uses the BioProv Model to answer provenance questions, such as “who was the collector of the specimens?”, “when and where they were collected?” and “who cataloged them?”.

- Time Ontology ([W3C, 2017](#)) is an ontology for temporal concepts developed by the World Wide Web Consortium (W3C). This ontology provides a vocabulary for expressing facts about topological relations among instants and intervals, together with information about duration and about date/time information.

In the next section, we will use all these ontologies to map biodiversity data into RDF triples.

4.3.2 The mapping component

The Mapping component of the STBioData architecture loads the domain ontologies, provenance model, taxonomic information and collection records to transform them in a set of RDF triples.

First, the collection records in CSV format are preprocessed, using the OpenRefine¹ program, to harmonize the names of species, the taxonomies and the name of *collectors* and *cataloguers*. OpenRefine is an open source tool from Google for working with noise data. For data available in ESRI shapefile format, STBioData uses the algorithm ([WARMERDAM, 2008](#)) to convert the ESRI shapefiles to WKT polygons in CSV. After that, provenance data is added using BioProv.

Now that all data are in the CSV format, STBioData uses a small mapping DSL (Domain Specific Language) ([AMANQUI et al., 2014](#)) to convert CSV data to RDF triples. These RDF triples are stored in a triple store (GraphDB²) and queried during user searches.

During the mapping process, the BioProv provenance model ([AMANQUI et al., 2016b](#)) is used to add provenance information. This model is based in the W3C PROV Data Model ([MOREAU; Missier \(Eds.\); W3C Provenance Working Group, 2013b](#)). BioProv defines a set of stating point terms divided into three classes: *Entity*, *Agents* and *Activities*.

In the BioProv model, the entity responsible for commanding the execution of an activity is modeled as an *Agent*. *Agents* are persons and organizations that deal with the biological collection information and interact with all activities and information that is updated. The *Collector* and *Cataloguer* agents (both *Agents* subclasses) are members of an organization.

A set of collected specimens is denominated *Collection Object* (CO) (an *Entity* subclass). A *CO* is generated by a *Collecting activity* (an *Activity* subclass). A *Collector agent* is associated with this activity. We trace the activity date using the *prov:startDateTime* property.

After a *collecting* process, the *CollectionObject* species needs to be identified through a *Cataloguing activity* (an *Activity* subclass). The *Cataloguer* agent assigns a unique identifier (taxon), to each *CollectionObject* using the taxonomic classification. The *Location* (an *Entity*

¹ <http://openrefine.org/>

² <http://java.icmc.usp.br:2190/sparql>

subclass) describes where the collection happened. It defines a locality: named place, habitat and spatial information (latitude and longitude).

These classes are associated by relations such as *prov:wasAttributedTo*, *prov:startedAtTime*, *bioprov:wasCollectedBy*, *bioprov:wasCatalogedBy*, *prov:actedOnBehalfOf*.

Figure 14 shows an example of BioProv provenance data. It was generated by the ProvValidator tool (MOREAU; HUYNH; MICHAELIDES, 2014) depicts relationships between agents, activities and locations.

The ProvValidator tool was used to validate BioProv data, generated from our biodiversity datasets, and to make sure it was encoded in standard interoperable RDF. The complete specification of the BioProv provenance model for biodiversity datasets is available online³.

4.3.3 STBioData web interface

The STBioData Web Interface is responsible for user interaction. It supports search queries for collection records using date ranges, linked biodiversity datasets visualization in an interactive map, provenance information visualization, data download in RDF or CSV formats and online dataset statistics.

Figure 15 shows the Web interface after a user (biodiversity specialist) began a search process with an initial keyword list (*Inia Geoffrensis*) and a date interval (1/1/1750 - 5/22/2017). The results display consists of following components (shown in Figure 15):

- (a) Search box, showing keyword list and data range;
- (b) Visualization of the matched collection records in an interactive map order. It shows quantity of specimens by geographic location of each record;
- (c) Visualization of provenance information for the selected collection record in the map. It includes collector and cataloguer identity and collection and cataloguing dates;
- (d) A vertical list of collection records that fulfill the search criteria showing information, such as family, genus and species.
- (e) Button to download the record set in CSV or RDF format;
- (f) Link to the visualization of dataset statistics .
- (g) Link to the SPARQL EndPoint.

Figure 16 shows the dataset statistics interface showing the occurrences quantity, species number, number of georeferenced occurrences, number of occurrences of Animalia Kingdom species, number of occurrences of Plantae Kingdom species, and a graphic showing the datasets were the records originated.

³ <https://www.researchgate.net/publication/301287278>

Figure 14 – Example of BioProv provenance data.

Source: Elaborated by the author.

Figure 16 – Statistics of biodiversity datasets stored in the STBioData prototype.

Source: Elaborated by the author.

The STBioData Web Interface was developed, using the Clojurescript language, as a Single Page Application (SPA). A SPA is a web application that fits on a single web page with the goal of providing a user experience similar to that of a desktop application (MIKOWSKI; POWELL, 2013). The Google Maps API was used to visualize georeferenced biodiversity data in an interactive map (DEVELOPERS, 2017). ClojureScript is a compiler for the Clojure language (a modern Lisp dialect) that targets JavaScript. It is designed to emit JavaScript code that is compatible with the advanced compilation mode of the Google Closure optimizing compiler (SIERRA; VANDERHART, 2016).

The queries, entered in the Web interface, are processed by the query reformulator component.

4.3.4 The query reformulator component

STBioData queries are more than simple keyword based search. Using the keywords presented by the user, the Query Reformulator finds ontology terms associated with them. It finds all ontology classes (taxonomic, spatial and temporal classes) to which the keywords belong and expands the original keyword list using the Algorithm 3. When a user submits a query, each word in the query is compared with the labels from the ontology classes in the GraphDB triple store (using SPARQL queries). If a concept also has WKT points with latitude and longitude data attached to it, these are also collected and showed in an interactive map. The algorithm results, obtained using multiple SPARQL queries, are presented to the user in the Web Interface.

Algorithm 3 – Algorithm to compare retrieve, link and visualize linked biodiversity data.

```

1: Input: userQuery a set of words typed by user
2: Output: results a information list from the triples
3: Connect to STBiodata triple store
4: for word in userQuery do
5:   submit a SPARQL query with word as subject
6:   add return to results                                ▷ word ?predicate ?object
7:   submit a SPARQL query with word as predicate
8:   add return to results                                ▷ ?subject word ?object
9:   submit a SPARQL query with word as object
10:  add return to results                                ▷ ?subject ?predicate word
11: end for
12: for unique objects and subjects in results do
13:   submit a SPARQL query to find its Geometry and WKT point location
14:   if Geometry = Null then
15:     submit a SPARQL query to find the WKT point of the County
16:   end if
17:   add return to results                                ▷ :Point, :longitude and :latitude
18: end for
19: for unique objects and subjects in results do
20:   submit a SPARQL query to find provenance information
21:   add return to results                                ▷ :NameCataloguer, :NameCollector, :DataCataloguing,
   :DataCollecting and :NameInstitution
22: end for
23: for unique objects and subjects in results do
24:   submit a SPARQL query to count the quantity of occurrences, species, occurrences by
   Animalia kingdom, occurrences by Plantae kingdom, Georeference occurrences
25:   add statistics to results ▷ :occurrences, :species, :Animalia, :Plantae and :Georeference
26: end for

```

This algorithm was implemented using the RDF4J (REST API) ([Eclipse RDF4J, 2017](#)), and GraphDB triple store.

4.4 Experiments

In order to validate STBioData architecture and guide the prototype tests, 10 biodiversity specialist, from the Laboratory of Molecular Biodiversity and Conservation (Federal University of São Carlos), were interviewed to define use cases with features and scenarios to identify the various user tasks. A list of our use cases is available online. In this chapter, we present one of these use cases:

- **USE CASE:** Determining the conservation status of *Inia geoffrensis* species from a date range
- **USER:** Jorge Luis, 31 years-old, PhD in Genetics and Evolution, university professor.

- GOAL: Obtain information about *Inia geoffrensis* species to define its conservation status.
- MOTIVATION: Reports about the conservation status of the *Inia geoffrensis* dolphin species support the process of identifying protected natural areas (inhabited by this dolphin). Preserving fragile habitats and ecosystems helps to ensure ecological balance and biological diversity in order to achieve sustainable use of natural resources.
- TASKS
 1. Specification of scientific name and date range;
 2. Find collection records for the *Inia geoffrensis* with geographical information (country, state, locality, latitude and longitude);
 3. Identify the name of the institutions where these collections are housed;
 4. Integrate the relevant collection records from biodiversity institutions (in this use case: INPA, EMBRAPA) using CSV spreadsheet data (collection code, catalog number, taxonomy information, scientific name author, collector name, etc);
 5. Integrate the resultant information, from step 4, with data from the IUCN Red List of Threatened Species to know the conservation status of the *Inia geoffrensis* species (in the year each collection was done);
 6. Use the information, from step 5, to plot a map. That was done, before the STBioData prototype was available, using tools such as the R Project (software environment for statistical computing and graphics) or other programs;
 7. Create a report with all the integrated information from step 5. This report is the basis to propose (or not) changes to the *Inia geoffrensis* conservation status.

Based in this use case, we decided to evaluate four aspects of the experimental prototype: creation of triples from heterogeneous sources (5.3.1); integrating existing data to determine the conservation status (5.3.2); use case testing: retrieving the conservation status of *Inia geoffrensis* species from a date range (5.3.3); and biodiversity information retrieval experiment (5.3.4).

4.4.1 *Creation of triples from heterogeneous sources*

The first experiment evaluated the linking mechanism between the Darwin Core, DBpedia, BioProv, GeoSPARQL and Time ontologies and the biodiversity data (from IUCN, EMBRAPA and INPA).

Using the SpeciesLink website, we downloaded 46568 records of animals and plants related to INPA and Embrapa institutions. All these records were in CSV format. Following this, we used a script, written in the STBioData mapping DSL (Section 5.2.2), to convert the

biodiversity data from INPA and Embrapa in RDF triples. The RDF dataset for each institution is available online⁴. We created, in total, 370,770 triples from INPA and Embrapa records.

Our second experiment evaluated the linking mechanism between Darwin Core, DBpedia, BioProv, GeoSPARQL and Time ontologies and IUCN. We downloaded marine mammals data from IUCN⁵ and read all marine mammals records in ESRI shapefile format (38,589 records in total) and transformed them to RDF triples (1,832,970 triples). The RDF dataset of IUCN marine mammals is available online⁶.

In both experiments we used the GeoSPARQL ontology to represent geometries and WKT format to represent latitude and longitude. We also used the Time Ontology to represent *collecting* and *cataloguing* dates and the BioProv to represent the provenance for each occurrence record. Terms from the DBpedia ontology were used to represent cities, states and countries. Finally, all the triples generated were loaded and made available in our GraphDB triple store.

Figure 16 shows the data metrics, the prototype loaded 46,568 occurrences, 38,588 occurrences of 9,216 *Animalia* species, and 7980 occurrences of *Plantae* (all georeferenced). Also, Figure 16 shows the quantity of occurrences by datasets. There are more occurrences from IUCN (82.6 %) because this institution store conservation status information from species around the world.

The triples, created in this experiment, were used to test the more advance search techniques shown in the next sections.

4.4.2 Integrating existing data to determine the conservation status

For this experiment, we used the triples generated from IUCN, EMBRAPA and INPA datasets created in Section 5.3.1. INPA and EMBRAPA datasets contain the occurrences of South American species, but do not contain information about conservation status. IUCN datasets contain the conservation status of species around the world and their geographical distribution. We linked these datasets to find the conservation status of *Inia Geoffrensis* in South America. STBioData integrated the datasets from INPA, EMBRAPA and IUCN following the four steps:

1. It read, using a SPARQL query, all the occurrences of *Inia Geoffrensis* in INPA and EMBRAPA datasets;
2. It read, using a SPARQL query, the conservation status of marine mammal species and their geographic ranges from IUCN datasets. Each species has a set of geographic ranges, one to every year with available data.

⁴ <http://java.icmc.usp.br:2190/graphs>

⁵ <http://www.iucnredlist.org/technical-documents/spatial-data>

⁶ <https://github.com/DilvanLab/STBioData>

3. For each species's geographic range, from step (ii), it added, to each range, collection data for the species and range year, from EMBRAPA and INPA datasets. In some cases, collection records from INPA and EMBRAPA had no spatial information. In these cases, geographic information from IUCN was used to improve the spatial information of these records.
4. It plotted this information in an interactive map.

Figure 17 shows the generated map. The circles (blue and yellow) group collections, by geographic areas, for the species and year range selected by the search parameters. The yellow circle marks the group with more collections. Clicking on the circles will expand them showing subgroups. The red rectangle shows the column Conservation Status added to the collection records of the *Inia geoffrensis* species. All this information can be downloaded in CSV or RDF format.

At the beginning of this experiment, INPA and EMBRAPA collection records, had no conservation status associated with them. After the integration process, they are now associated with the conservation status for the species in the year the collection took place. Also, all have some form of georeferenced location.

Figure 17 – Web interface of integrating existing data to determine the conservation status.

Source: Elaborated by the author.

One of the big advantages of having biodiversity data in RDF is to be able to connect different sources. We created a SPARQL query to integrate the two different datasets: EMBRAPA and INPA collection dataset and the IUCN conservation status dataset.

Source Code 7 show a SPARQL query used to retrieve collection records and conservation status of the *Inia geoffrensis* dolphin from the 1987 to 2016 date range (Lines 28 to 29) collected inside of a Brazilian protected area (Lines 9, 18 to 19). Lines 1 to 5 select the variables to be returned by the query. Lines 6 to 17 specify the relationships between the values (scientific name, institutionCode, WKT points, data collected, name of the collector). Lines 23 to 27 select the conservation status from IUCN.

Source code 7 – SPARQL query used to retrieve collection records and conservation status of the *Inia geoffrensis* dolphin from the year 1987 to 2016.

```

1: Select ?id ?scientificname ?institutionCode
2: ?wktA ?dateCollected ?nameCollected where {
3: ?id dwc:institutionCode ?institutionCode.
4: ?id geo:hasGeometry/geo:asWKT ?wktA.
5: ?id prov:hadMember/bioprov:wasCollectedBy/
6: prov:startedAtTime ?dateCollected.
7: ?id prov:hadMember/bioprov:wasCollectedBy/
8: prov:wasAssociatedWith/foaf:givenName ?nameCollected.
9: ?id prov:hadMember/bioprov:wasCataloguingBy/
10: prov:wasAssociatedWith/foaf:givenName ?nameCataloguing.
11: FILTER (geof:ehInside(?wktA,
12: Polygon(-1 -58, -7 -58, -7 -69, -1 -69, -1 -58)))
13: OPTIONAL {
14: ?id bioprov:hasConservationStatus/dwc:occurrenceStatus
15: ?occurrenceStatusIUCN1.
16: ?id dwc:occurrenceStatus ?occurrenceStatusIUCN2. }
17: BIND(COALESCE(?occurrenceStatusIUCN1,
18: ?occurrenceStatusIUCN2) as ?occurrenceStatusIUCN)
19: FILTER (?dateCollected >= "1987")
20: FILTER (?dateCollected <= "2016")}
```

The query showed in Source Code 7 retrieves the same kind of information of the STBioData Web Interface in a textual way. It is not suitable to be used by end users, but, such queries, could be used by third part systems using the STBioData SPARQL End Point⁷. It also shows that, once data are converted to Linked Data format, it is easy to join different data sources. In the next section we show an exploratory evaluation of the use case.

⁷ <http://java.icmc.usp.br:2190/>

4.4.3 Use case testing: retrieving the conservation status of *Inia geoffrensis* species from a date range

For this experiment, 10 biodiversity domain experts from the Laboratory of Molecular Biodiversity and Conservation⁸ used the STBioData prototype in the use case defined in the beginning of this section: Determine the conservation status of the *Inia geoffrensis* species in different time ranges. The domain experts were asked to execute the following tasks:

1. Specify scientific name and date ranges;
2. Find geographic information of *Inia geoffrensis* collection records;
3. Identify the collection code, catalog number, taxonomy information of *Inia geoffrensis* specimens;
4. Check the conservation status for *Inia geoffrensis* species in each collection;
5. Visualize the collection information in an interactive map
6. Export the results as a report in CSV format and print the map in PDF format.

These tasks are a simplification of the 7 tasks defined for this use case (Section 5.3). Parts of the original tasks are now automatically performed by the program (such as integration of records from different institutions). After finishing the tasks, the domain experts answered a questionnaire with in order to know their feedbacks and suggestions.

After finishing the tasks, the domain experts answered an evaluation questionnaire questionnaire (available online<https://www.researchgate.net/publication/317049932_Exploratory_evaluation_through_questionnaires>). It had two parts. The first part, with five questions, was used to determine the research area, scientific interests and software tools used by the participant. The second, with nine questions, is a subjective self-assessment of the completed tasks. With this questionnaire, we assessed the expert's experience in biodiversity and his positive and negative feedback about aspects of the STBioData prototype.

The answers of the first part of the questionnaire shows that 90% of the experts work with biological sciences and 10% with agricultural sciences. 100% of them have research interest in Conservation Genetics. Figure 18 shows the most common search tools used by the biodiversity experts to search for species collection records.

Google Scholar and Web of Science are the most used repositories, with 7 points. The IUCN site has 4 points, SpeciesLink, Cielo e NCBI have 3 points, Google has 2 and Arkive has only 1. 70% of the biodiversity experts are doing data integration manually. For the generation of

⁸ <http://www.dge.ufscar.br/>

Figure 18 – The most common Keyword search tools used by the specialist in biodiversity to search occurrences of species.

Source: Elaborated by the author.

geographic visualizations, they are using tools like QGIS⁹, DIVA-GIS¹⁰, EARTH EXPLORER¹¹, IBGE¹², MMA-BIOMAS¹³ biomas to visualize collection data in a map.

The answers to the second part of the questionnaire showed that 90% of the participants found information for other dolphin species (besides *Inia geoffrensis*) and were satisfied with the recovery time. 90% of the experts did not have problems to understand the interactive map. 10% said that it was necessary to explain the yellow circle in the map (the yellow circle shows the set with more collected specimens). 100% of the experts were able to describe the purpose of the tool. Users also made good recommendations, such as “In the SpeciesLink system we cannot see collection quantities in a map, but, in STBioData, we can see how many specimens collection are available in each geographic location”. They recommended improvements to our architecture with the addition of more repositories, such as NCBI¹⁴, Web of Science¹⁵, Fishbase¹⁶, Scielo¹⁷, GenBank¹⁸, and PubMed¹⁹.

After this exploratory evaluation, we found out that, according to the experts, (i) the quality of the final data is better when using STBioData instead of SpeciesLink, by using the

⁹ <http://www.qgis.org/es/site/>

¹⁰ <http://www.diva-gis.org/>

¹¹ <https://earthexplorer.usgs.gov/>

¹² <https://www.ibge.gov.br/>

¹³ <http://www.mma.gov.br/>

¹⁴ <https://www.ncbi.nlm.nih.gov/>

¹⁵ <https://webofknowledge.com/>

¹⁶ <http://www.fishbase.org/>

¹⁷ <http://www.scielo.org/php/index.php>

¹⁸ <https://www.ncbi.nlm.nih.gov/genbank/>

¹⁹ <https://www.ncbi.nlm.nih.gov/pubmed/>

map and integrated information downloaded in CSV format; (ii) showing, in the interactive map, the quantity of collections grouped by geographical location allow users to more easily estimate the quantity of individuals in a region. This information is very important when making any decision about an area.

4.4.4 Biodiversity information retrieval experiment

For this experiment, a group of 10 biodiversity domain experts, from the Laboratory of Molecular Biodiversity and Conservation, which work with marine mammal species, were asked to create two different queries to find collection records for the species they work with. They had to run these two queries in the STBioData prototype and the SpeciesLink site. The STBioData prototype was loaded with all records about marine mammals from SpeciesLink that came from INPA and EMBRAPA. We ended up with 20 queries for STBioData and 20 for SpeciesLink.

After we tested each query, each expert was asked to judge which results were relevant and non relevant. We compared the results from the two search systems: the STBioData semantic search and SpeciesLink keyword based search. Figure 19 shows the evaluation of the two systems, using the average precision, recall and F1-measure.

Figure 19 – Precision and recall experimental results.

Source: Elaborated by the author.

The STBioData semantic search was, on average, 24% more precise than the SpeciesLink keyword search, in the queries examples from the domain experts. The recall in STBioData was 22% better than SpeciesLink and the F1-measure was 28% better. The results for STBioData were significantly better.

We extended these measures to evaluate the ranked retrieval results for STBioData and SpeciesLink searches. The traditional way of doing this is the 11-point interpolated average precision. For each such set, precision and recall values can be plotted to give a precision-recall curve considering the top $k=10$ retrieved documents (Figure 20).

Figure 20 – Averaged 11-point precision/recall graph across 20 queries for a representative STBioData and SpeciesLink search system.

Source: Elaborated by the author.

It is very important to point out that the data, shown in Figure 20, only takes into account cases where the SpeciesLink keyword based search actually returned something. If the keywords are not present in the records, methods based only on keyword search fail. As an example, if users try to find records of the species "*Inia Geoffrensis*" between 2001 and 2005 in the SpeciesLink site, they will find none. The problem is that the SpeciesLink site does not consider a specific date range to search species. Using the STBioData to search for "*Inia Geoffrensis*" using the same time period would result in 80 records.

4.5 Conclusions and future work

The STBioData architecture is the main contribution of this paper. It automatically links spatiotemporal biodiversity data, from heterogeneous data sources, to enable easy data searching, visualization and downloading. We implemented and tested a prototype based on it. This prototype has a web interface, to be used by biodiversity domain experts, and a SPARQL End Point, to be used by third part software. For the testes, 46,121 collection records from the

SpeciesLink site and 38,589 conservation status records from the IUCN site were converted to 2,233,782 RDF triples and linked using well known ontologies.

During the tests, 90% of biodiversity domain experts using the tool to determine the conservation status were able to find information about dolphin species, were satisfied with recovery time and were able to understand the interactive map. In the information retrieval experiment, when compared with SpeciesLink search, the prototype's semantic based search performed, on average, 24% better in precision and 22% in recall tests. And that takes into account only searches where the SpeciesLink keyword based search actually returned a result.

Another important goal for us is to demonstrate the value of having linked biodiversity data public available in RDF format (or other Semantic Web standard format). The semantic information embedded in the data, using well known ontologies, allows the automatic integration of data from different independent sources. For an interdisciplinary field, such as biodiversity, that is a very interesting feature.

As future work, we plan to refine our experiments with other use cases in partnership with our biodiversity partners. There are different aspects of the proposed architecture that need to be assessed in future work, such as the performance in complex SPARQL and GeoSPARQL queries and the inclusion of more data sources.

The BioProv provenance model can be used to partially recover missing or incomplete biodiversity information. This could help to improve biodiversity data quality.

The search process, used in STBioData, can be optimized including (i) the option to choose species names navigating a taxonomic classification, (ii) the option to search data from specific areas on the map (chosen with a few mouse clicks), and (iv) the option to select interesting geospatial areas, such as biomes and natural parks, in the search interface.

4.6 Remarks about the chapter

In this chapter, we investigated our main research question, “How can we integrate heterogeneous biodiversity semantic data using their spatiotemporal information and provenance?”, by implementing and testing the STBioData architecture. STBioData automatically link spatiotemporal biodiversity data, from heterogeneous data sources, to enable easier searching, visualization and downloading of relevant data. We built a web based prototype: the STBioData system that implements the STBioData architecture.

The prototype was tested using a case study: determining the conservation status of *Inia Geoffrensis*, an Amazonian River Dolphin. For the tests, 46,211 collection records, from GBIF, and 38,589 conservation status records (including maps), from IUCN, for marine mammal were converted to 2,233,782 RDF triples and linked using well known ontologies.

Users can access the STBioData system, using the web interface, and search for collection

and distribution records based on species names and time ranges. When testing the system, 90% of biodiversity domain experts were able to find information about the dolphin species they worked with (to determine their conservation status), were satisfied with the recovery time and were able to understand the interactive map. In the information retrieval experiment, when compared with SpeciesLink search, the prototype's semantic based search performed, on average, 24% better in precision and 22% in recall tests. And that takes into account only searches where the SpeciesLink keyword based search actually returned a result.

BioProv enables STBioData to incorporate provenance data. This has the potential to answer questions, such as: Who was the specimens collector? Who was the cataloguer of the species? When was the data collected? But more important, provenance data can be used to link data and answer questions like, what other specimens were collected in the same place? Which species a cataloguer works with? Or which specimens were collected in a particular protected area?. We created a SPARQL query used to retrieve provenance information from collection records and conservation status of the *Inia geoffrensis dolphin* from the 1987 to 2016 date range collected inside of a Brazilian protected area.

We conducted quantitative and qualitative evaluations of the STBioData system to find out if it was able to help biodiversity experts retrieve and link biodiversity information. The system was evaluated by biodiversity experts from the Laboratory of Molecular Biodiversity and Conservation (UFScar), and by Semantic Web specialists from the University of São Paulo and the Data Science Laboratory (from the University of Ghent - Belgium). 90% of the experts did not have problems to understand the interactive map of STBioData. 10% said that it was necessary to explain the yellow circle in the map (the yellow circle shows the set with more collected specimens). 100% of the experts were able to describe the purpose of the tool.

In the next chapter, we will present an extension of the STBioData architecture to answer more advanced geospatial queries.

LINKING BIODIVERSITY AND GEOSPATIAL DATA

In this chapter, the research question 4, "How can we improve the location accuracy of biodiversity data?", is investigated. The chapter article, entitled "An Architecture to Create Geospatial Queries for the Biodiversity Domain, a Use Case about Endangered Species in Brazilian Biomes", proposes an extension of the STBioData architecture to answer more advanced geospatial queries. The first version of STBioData, described in chapter 4, supports searches based on species names and date ranges. However, it does not allow for searches of implicit relationships, such as spatial relationships. For example "Which occurrences of a species are inside of the Atlantic and Amazon Forests?". To solve this need, this chapter describes an architecture to link biodiversity and geospatial information, from heterogeneous data sources, to enable easier searching, visualizing and downloading of relevant data.

In the chapter 4, we used the GeoSPARQL ontology to represent the spatial data and the Time ontology to represent the temporal data of biodiversity collections. We will reuse these representations to test more advanced searches; for example, the user can search for occurrences of species inside of biomes and natural parks.

This chapter is organized as follows. Section 5.1 describes the motivation behind the extension of STBioData architecture. The extension of STBioData architecture is presented in Section 5.2. Section 5.3 presents our experiments and use case. Section 5.4 describes the related work. Section 5.5, concludes by summarizing our results and describing future work. Finally, the remarks of this chapter are given in Section 5.6.

5.1 Introduction

Biodiversity data is increasingly being made available on the Web in the form of datasets described using the Resource Description Framework (RDF) ([LASSILA *et al.*, 1998](#)). RDF

exposes data as triples, consisting of subject, predicate and object, stating that subject is related to object by the predicate relation. The principles of Linked Open Data (LOD) ([BERNERS-LEE, 2006](#)) encourage a set of best practices for publishing and connecting structured data on the Web. Linked Open Data promotes the use of the SPARQL Protocol and RDF Query Language (SPARQL) ([PRUD'HOMMEAUX; SEABORNE, 2006](#)) and RDF to query and model data. While this is useful for querying for relationships that are explicitly represented in data, implicit relationships, such as spatial relationships, cannot easily be queried. For instance, datasets may exist that describe protected areas, biomes and natural parks in Brazil, but being able to link these datasets based on their undeclared relationships is difficult. The ability to answer a meaningful query, for example like “What occurrences of species are inside of *Atlantic* and *Amazon* Forest?”, depends on how the data is represented, whether all of the resources are related to the *Atlantic* and *Amazon* Forest area, and if that relationship is explicit.

To further clarify this problem, we use the example of finding species inside of *Atlantic* and *Amazon* Forest. Data related to occurrences of species are collected and saved in a csv file by *Collectors*. After a *cataloging* process, the data are confirmed to belong to specimens of species by *Cataloguers*. This information is published in web sites like SpeciesLink. This file could be in csv format and contain information about the location of the species occurrence. Another researcher generates the polygon of the *Atlantic* and *Amazon* forest area and publish these information in the format of ESRI shapefiles. Then, a biodiversity specialist wants to integrate all the previous information (occurrences of species from SpeciesLink and polygons of the *Atlantic* and *Amazon* forest) to create a report to know What are the occurrences of species inside of *Atlantic* and *Amazon* Forest area. In other words, he/she wants to integrate the information from heterogeneous sources. This is a common problem in the biodiversity domain.

To help solve this problem, we present an architecture to link biodiversity and geospatial information, from heterogeneous data sources, to enable easier searching, visualizing and downloading of relevant data. It is an extension of the STBioData architecture ([AMANQUI; POTENCIANO; MOREIRA, 2017b](#)).

STBioData automatically link spatiotemporal biodiversity data from heterogeneous data sources. The extension adds the following novelties:

- The search process now uses GeoSPARQL functions (within, intersects) in Brazilian *Biomes* and/or *Natural Parks*.
- Two new online metrics were implemented: occurrences in South America and occurrences organized by collection year (these features were requests from biodiversity experts).
- The option of seeing the provenance data for each specimen collection in an interactive map.

A new use case scenario was introduced that better reflects a realistic example of how STBioData can be used by biodiversity specialists. The new STBioData architecture was tested using a representative biodiversity dataset from the National Institute for Amazonian Research (INPA), the Brazilian Agricultural Research Corporation (EMBRAPA) and the IUCN Red List of Threatened SpeciesTM(IUCN). A prototype system was evaluated by experts from the Laboratory of Molecular Biodiversity and Conservation from the Federal University of São Carlos (UFScar).

We conducted an exploratory evaluation, using questionnaires, to find out how the architecture helped biodiversity specialists retrieve and link geospatial information. Our results show that Cerrado and Atlantic Forest were the subjects that our participants showed the most interest in. 100% of the participants found the information about of the species that they were searching and they were satisfied with the amount of recovery time. 100% of the specialists did not have problems understanding the interactive map and they were able to describe the purpose of the new version of STBioData.

The remainder of this article proceeds as follows: Section 5.2 shows the extension of STBioData Architecture. Section 5.3 presents our experiments and use case. Section 5.4 discusses the related work. Finally, Section 5.5 concludes by summarizing our results and describing future works.

5.2 Extension of STBioData architecture

Figure 21 presents the new STBioData extended architecture for linked biodiversity and geospatial data. It uses GeoSPARQL functions (within, intersects, etc.) to link occurrences of species (collections) inside Brazilian biomes and/or natural parks. We expanded the search process in STBioData, added the option of using spatial functions and to select places of interest (Brazilian biomes or natural park). We also added an option to visualize provenance in an interactive map. STBioData now also show polygon drawings of biomes and natural parks in the interactive map, according to the query results. Collections can be organized by year collected.

As shown in Figure 21, the STBioData architecture is formed by layers. The Presentation layer has a web interface that enables user interaction, using a user interface with digital maps, and a GeoSPARQL endpoint for software agent access. In this way, it is possible to support both human and machine interactions for inserting or finding any information about specimen collection in the STBioData dataset. The Business logic layer processes the information retrieval through the query builder component. This new component will be explained in the next subsection. The Data layer provides access to the RDF triples stored in the Triple Store (in the prototype, the GraphDB store¹). The Model layer contains a collection of concepts and relationships that characterize the knowledge in the biodiversity domain (domain ontologies).

¹ <http://java.icmc.usp.br:2190/sparql>

Figure 21 – STBioData architecture extended to link geospatial data.

Source: Elaborated by the author.

5.2.1 Web interface

The Web interface is shown in Figure 22 and is responsible for the interaction between users and our architecture. It supports (i) search queries within a date range; (ii) GeoSPARQL functions (within and intersects); (iii) visualization of linked biodiversity datasets in an interactive map; (iv) visualization of provenance information; (v) data export function using RDF or CSV formats and (vi) online dataset metrics (statistics).

An interactive search process begins with an initial keyword list entered by a user (a biodiversity expert) that represents his/her search intentions. The name of the species can be defined in the Feature 1 panel (query area surrounded by an orange rectangle in Figure 22). A spatial GeoSPARQL function can be defined in the Feature 2 panel (blue rectangle in Figure 22). The Feature 3 panel can be used to select one or two places of interest (e.g. biomes and natural parks) (red rectangle in Figure 22). The Feature 4 panel is used to select a date range. (yellow rectangle in Figure 22).

Source: Elaborated by the author.

The search results display consists of the records visualization in an interactive map and table. [Figure 22](#) shows the interactive map with geographic features (places and points of interest). The interface also has the option of showing the generated GeoSPARQL query used to fetch data (black rectangle in [Figure 22](#)).

A vertical list of document titles and several lines of records that fulfill the search criteria (family, genus, species and other information) is shown in [Figure 22](#) (purple rectangle). These records can be downloaded in CSV format or RDF format.

The visualization of the provenance information is shown in [Figure 23](#). In STBioData,

Figure 23 – Visualization of the provenance information in STBioData.

Source: Elaborated by the author.

users can also see two new online statistics: occurrences in South America and occurrences organized by collection year. These two statistics were developed following requirements from biodiversity experts. [Figure 24](#) shows occurrences in South America, yellow means less and green more. In this case, most occurrences were in Brazil.

The occurrences by year collected are presented in [Figure 25](#).

5.2.2 The query builder component

We added a new component to the business layer to process search strings in natural language, sent by the Presentation Layer. The query builder component receives the search terms from users, finds all ontology classes to which the term belongs and expands the keyword list using [Algorithm 4](#).

Figure 24 – Occurrences in South America.

Source: Elaborated by the author.

Figure 25 – Occurrences organized by year collected.

Source: Elaborated by the author.

When a user submits a query (selecting the species name, date range, geospatial function and biomes or natural parks) each value is compared with the labels from the triples in the GraphDB triple store (using GeoSPARQL queries). If the triples also have WKT points with latitude and longitude data attached to them, these are also collected and compared with the polygon area of the biomes or natural parks, and showed in the interactive map.

The STBioData prototype was implemented as a Single Page Application (SPA), using Clojurescript ([SIERRA; VANDERHART, 2016](#)) (a new Lisp language targeting web applications). The server side uses the RDF4J (REST) API ([Eclipse RDF4J, 2017](#)) to reason about

GeoSPARQL queries, in addition to a GraphDB triple store. To visualize the geofenced biodiversity data, first we process the data, with Clojurescript, and then send it to the Google Maps API (DEVELOPERS, 2017) that print the polygons and points inside the map. The source code for the STBioData implementation can be found in the GitHub repository².

Algorithm 4 – Query builder algorithm to compare retrieve, link and visualize linked biodiversity and geospatial data.

```

1: Input: userQuery a set of words typed by user
2: Output: results a information list from the triples
3: Output: SPARQLquery a SPARQL query to execute in other triples stores
4: Connect to STBiodata triple store
5: for word in userQuery do
6:   submit a SPARQL query with word as subject
7:   add return to results                                     ▷ word ?predicate ?object
8:   submit a SPARQL query with word as predicate
9:   add return to results                                     ▷ ?subject word ?object
10:  submit a SPARQL query with word as object
11:  add return to results                                     ▷ ?subject ?predicate word
12: end for
13: for unique objects and subjects in results do
14:   get Geometry, GeoSPARQLfunction and PolygonBiome
15:   submit a SPARQL query to find its Geometry and WKT point location
16:   if Geometry = Null then
17:     submit a SPARQL query to find the WKT point of the County
18:   end if
19:   add return to results                                     ▷ :Point, :longitude and :latitude
20: end for
21: for unique objects and subjects in PolygonBiome do
22:   submit a query to find its Geometry and WKT point location using the
   GeoSPARQLfunction and PolygonBiome
23:   save SPARQLquery
24:   if PolygonBiome = Null then
25:     submit a SPARQL query without GeoSPARQL function
26:     save SPARQLquery
27:   end if
28:   add return to results                                     ▷ :Point, :longitude, :latitude, PolygonBiome
29: end for
30: for unique objects and subjects in results do
31:   submit a SPARQL query to find provenance information
32:   add return to results                                     ▷ :NameCataloguer, :NameCollector, :DataCataloguing,
   :DataCollecting and :NameInstitution
33: end for

```

² <https://github.com/DilvanLab/STBioData>

5.3 Experiments

In order to validate our architecture and guide the prototype tests, six biodiversity experts, from the Laboratory of Molecular Biodiversity and Conservation (UFSCar), were interviewed to define a set of queries considering biomes and natural parks of Brazil. They also helped to create a use case to use the new map visualization showing the species occurrences inside biomes and natural parks:

- **USE CASE:** Create a report of occurrences of species inside of Brazilian *biomes and natural parks*
- **USER:** Karla Verónica Chávez Rodríguez de Congrains , 31 years-old, Master en Ecology, Federal University of São Carlos.
- **GOAL:** Identifying species from Brazilian Biomes
- **MOTIVATION:** Reports of species occurrences inside biomes and natural parks are an important tool for conserving species and ecosystems.
- **TASKS**
 1. Find the information of polygons of Brazilian Biomes and Natural Parks in the Chico Mendes Institute (in ESRI shapefiles);
 2. Find geographic information of species occurrences in the SpeciesLink web site (in CSV format);
 3. Using tools such as R, QGis, Google Earth to visualize and find specimens inside of biomes and natural parks from step 1;
 4. Use tools, such as R, QGis and Google Earth, to visualize and find species occurrences inside the biomes and natural parks (from step 1);
 5. Use tools, such as R, QGis, Google Earth, to plot maps of Brazilian Biomes and geographic locations of species occurrences from SpeciesLink;
 6. Use the results to generate reports.

5.3.1 GeoSpatial dataset

To create a report of occurrences of specimens inside of Brazilian biomes and natural parks. We first needed to find the polygons of natural parks and biomes in Brazil and convert them in a format that could be used by SPARQL EndPoints.

According to our interviews, the biodiversity specialists recommended us to download the ESRI shapefile of biomes or natural parks from the Chico Mendes Institute for Biodiversity

Conservation (ICMBio)³. ICMBio is an important institution for protecting Brazil's natural heritage and promoting biodiversity conservation.

We downloaded 09 ESRI shapefiles from ICMBio about the following biomes and natural parks:

- *Amazon and Atlantic Forest*;
- *Caatinga, Cerrado, Mata Atlantic, Pampa and Pantanal* biomes;
- *Jaú, Mapinguari and Pico da Neblina National Park*.

We used the QGIS to open and extract the complete polygon of each ESRI shapefile. But when we executed the SPARQL query with these polygons, our EndPoint SPARQL did not shown results because there were to many points in each polygon. For that reason, biodiversity specialist advised us to reduce the quantity of points of the polygons. We used the option “Polygon from layer extend” of the QGIS system to reduce the quantity of points of each polygon.

The complete list of polygons are published online⁴. These polygons were validated for the biodiversity specialists. We used biodiversity triples from INPA and EMBRAPA generated in chapter 4. These triples and the polygons allow us to implement more advanced search methods for querying species occurrences.

These tests are shown in the next section.

5.3.2 User Queries

For the previous use case, 06 biodiversity specialists from the Laboratory of Molecular Biodiversity and Conservation⁵ of the Federal University of Sao Carlos created query examples that should have returned occurrences of species inside of Brazilian *biomes and natural parks*. A total of 10 distinct queries were done. The list of queries that we used are shown below:

1. Find *Phyllomys dasythrix* the occurrences of species that intersects Pampa Biome and Atlantic Forest between 1950 and 2016
2. Find *Reithrodon typicus* occurrences of species within *Pampa Biome* between 2005 and 2010
3. Find *Bacopa monnierioides* occurrences of species within *Pantanal Biome and Cerrado Biome* between 1980 and 2000

³ <http://www.icmbio.gov.br/portal/geoprocessamentos>

⁴ <https://github.com/DilvanLab/STBioData/>

⁵ <http://www.dge.ufscar.br/>

4. Find all the occurrences of species that intersects *Caatinga and Atlantic Forest* between 1980 and 1985
5. Find *Tapirus terrestris* occurrences of species within *Mapinguari National Park* between 2013 and 2016
6. Find all the occurrences of species that intersects *Jaú National Park* between 2007 and 2012
7. Find *Inia geoffrensis* occurrences of species within *Amazon Forest*
8. Find *Cardiospermum anomalum* occurrences of species that intersects *Cerrado and Caatinga Biome* between 1980 and 1985
9. Find *Akodon montensis* occurrences of species within *Atlantic Forest* between 2000 and 2010
10. Find all occurrences of species within *Amazon Forest*.

Our first idea was compare the results of two search systems: our STBioData and SPARQL EndPoint, both using data from INPA and EMBRAPA. However, when the biodiversity specialists started the search process for the first query in our EndPoint SPARQL, they took a long time to understand the syntaxes of a SPARQL query. We explained the basic concepts of SPARQL queries to each biodiversity specialist but we showed that is difficult for them create and modify a SPARQL query. For that reason we decided to compare the amount of time for the specialists to answer the 10 queries in two different system: the STBioData System and the SpeciesLink system.

STBioData and SpeciesLink have the option of showing the natural parks and biomes in a map and the information can be downloaded in a CSV file. 06 biodiversity specialists executed 10 queries in both systems and we compared the amount of time it took each biodiversity specialist to complete the search process in the STBioData and SpeciesLink. We timed each biodiversity specialist from the start of the search process and ending when he/she press the CreateQuery button in STBioData. The same in SpeciesLink, we measured the time from the start of the search process until they press the “Buscar” button.

Figure 26 shows the generated map for a query in the STBioData system. This map visualizes all the occurrence of specimens that intersects Amazon Forest and Pico da Neblina National Park. STBioData found 35 occurrence of specimens. To compare STBioData and SpeciesLink systems, we used the metrics: number of query, name of the system and amount of time in seconds to answer the query. This configuration was executed for 06 users and for 10 queries.

The configurations of each query was executed 6 times with a confidence of 95% according to a distribution t-student. The results of the p-values were down to 0.05, all the results

Figure 26 – Map for a query in STBioData system. This map visualizes all the occurrence of specimens that intersects Amazon Forest and Pico da Neblina National Park. STBioData found 35 occurrence of specimens.

Source: Elaborated by the author.

Figure 27 – Evaluation of STBioData and SpeciesLink keyword search system.

Source: Elaborated by the author.

of the comparison shown in the Figure 27 are different statistics. Figure 28 shows the evaluation of two systems using the amount of time in seconds to answer the 10 queries. In this figure we can see that each biodiversity specialist took less time to do the search process in the STBioData.

The STBioData system also has the option to generate a SPARQL query in the search process. The complete list of SPARQL queries generated in this experiment were published online⁶.

⁶ <https://github.com/DilvanLab/STBioData>

Figure 28 – Evaluation of the two systems using the amount of time in seconds to answer the 10 queries.

Source: Elaborated by the author.

In Source Code 8, we show a full SPARQL query that uses GeoSPARQL functions to recover occurrences of the species *Bacopa monnierioides* within *Pantanal* and *Cerrado Biome* between 1980 and 2000 :

Source code 8 – SPARQL query that use GeoSPARQL functions to recover occurrences of the species *Bacopa monnierioides* within *Pantanal* and *Cerrado Biome* between 1980 and 2000.

```

1: SELECT ?CodSpecie ?scientificname ?pontowkt ?dateCollected
2: WHERE {
3: ?CodSpecie geo:hasGeometry/geo:asWKT ?pontowkt .
4: ?CodSpecie dwc:scientificName ?scientificname .
5: ?CodSpecie prov:hadMember/bioprov:wasCollectedBy/prov:
    startedAtTime
6: ?dateCollected .
7: FILTER (?scientificname = "Bacopa monnierioides").
8: FILTER (geof:sfWithin(?pontowkt , "Polygon((-15.3040028213148
9: -15.7444231432591 -54.3147335282668, -21.2006133574037...
    -59.3549951567885))
10: "^^ogc:wktLiteral))
11: FILTER (geof:sfWithin(?pontowkt , "Polygon((-1.78245627439619
    -43.7797486666813,
12: -53.3444687732311, ...
13: -1.78245627439619 -43.7797486666813))"^^ogc:wktLiteral))
14: FILTER (?dateCollected >= "1980")
15: FILTER (?dateCollected <= "2000")}
```

In order to validate our experiments we conducted a qualitative evaluation of STBioData by presenting to the specialist in biodiversity one questionnaire. This evaluation is showed in the next subsection.

5.3.3 Qualitative evaluation

We conducted an exploratory evaluation of our STBioData by presenting to the specialist in biodiversity one questionnaire. The model of these questionnaires is available online⁷.

The purpose of the first part of the questionnaire was to determine the research area and scientific interests of the participants. The purpose of the second part of the questionnaire allowed for a subjective self-assessment of the previously completed use case. With this questionnaire we assessed the specialist in biodiversity's experience with our architecture and determined what the positive and negative aspects of STBioData are.

Of the people invited, 66.7% of the participants have a research background in biological sciences and 33.3% of participants have a research background in exact and earth sciences. The average time taken for the evaluation was 24 min 17 s.

Cerrado and Atlantic Forest were the subjects that our participants showed the most interest in. 100% of the participants found the information about of the species that they were searching and they were satisfied with the amount of recovery time. 100% of the specialists did not have problems understanding the interactive map and they were able to describe the purpose of STBioData.

Another interesting part of the results were the comments of the participants. We also grouped the suggestions of the biodiversity specialists into two categories: suggestions for improving the search process and suggestions for improving the visualization of the results. These two categories are listed below:

- Search Process: the taxonomy of the species is important in the moment of the search process. (e.g., *"Sometimes I do not remember the name of the species, I prefer to search the species name by a taxonomy"*) Improve the search in range of data. (e.g., *"I would like to search only for year of collected data and not for all data collected"*).
- Visualization of the Results: the biodiversity specialists wanted to export the map generated in geospatial formats and also add the option of searching in other natural parks (e.g., *"I want to add other natural parks in the search process and export the map generated in KML"*). Generate the map in a shorter amount of time (e.g., *"When a user wants to generate a map that intersect two polygons the STBioData take more time to process this query than the others"*).

⁷ <https://docs.google.com/forms/d/1to8PTNs3NMDNjV2ew1CanPFBBY3KXLyVB2BrQgrtZdU/prefill>

5.4 Related work

In this work, we extended the STBioData architecture ([AMANQUI; POTENCIANO; MOREIRA, 2017b](#)). We added the possibility of enabling easy searching, visualizing and downloading of relevant data regarding geospatial areas (biomes and natural parks) and functions of GeoSPARQL. It supports the generation of interactive maps describing this data. An extension of the STBioData prototype based on this architecture was implemented. It was tested using a representative datasets about biodiversity and it was evaluated by biodiversity specialists.

We conducted a systematic review in the literature for works that dealt with linked geospatial and biodiversity data. The articles that we found only use biodiversity data or geospatial data, but not both. Below we list them:

Ying Zhang et al ([ZHANG *et al.*, 2016](#)) proposed a geospatial data retrieval model: Karma. Karma supports extracting geospatial data from CSV and RDF sources and geographic information visualizing on the map. The authors integrated data from OpenStreetMap, Wikimapia and GNIS data sources as examples. Our approach also considers using a map to visualize the geospatial data and also showing the provenance of this data.

Stadler et al ([STADLER; MARTIN; AUER, 2014](#)) proposed a Web browser for SPARQL accessible data with geospatial capabilities. This Web browser retrieves information from a set of related resources and shows this data on a map.

Lascata et al. ([LACASTA *et al.*, 2014](#)) created a Linked Geospatial Data model describing the Web Map Services published by Spanish organizations. This model allows semantic searching and browsing. This is done by crawling the web in search of available geospatial services and enriching their descriptions with concepts from common knowledge organizations' models. However, these approaches do not help users to search, visualize and download resultant data. Also they do not consider to use a provenance model that help to find the lineage data.

In ([AMANQUI *et al.*, 2014](#)), we proposed a semantic search architecture that uses ontologies and vocabularies to retrieve the biodiversity information. This approach is good for biodiversity data without geographical location, but in the cases when a user wants to recover data from a specific area, this is still a big issue.

Our challenges, which we tackle in STBioData, are linking biodiversity and geospatial data to answer complex queries, i.e., queries that use spatial relations (“what occurrence of species are inside of two protected areas?”). The related work detailed above can answer SPARQL queries, but they cannot handle queries with spatial concepts (such as occurrence of species within the *Amazon* and *Atlantic* Forests in a range of dates). To deal with this, STBioData uses ontologies to find new relations between places and perform spatially complex queries using the GeoSPARQL functions. Finally, to deal the biodiversity provenance information, STBioData uses the BioProv model.

5.5 Conclusion and future work

In this chapter, we presented an extension of STBioData for linking biodiversity and geospatial data. The proposed STBioData extension adds the option of selecting GeoSPARQL functions (within, intersects) and of selecting biomes and natural parks in the search process. Two new online metrics were implemented: occurrences in South America and occurrences organized by year collected. These two statistics were developed according to requirements of the domain experts after a interview. We also implemented the option of seeing the provenance of each occurrence of species in an interactive map.

STBioData was evaluated by users from Laboratory of Molecular Biodiversity and Conservation of the Federal University of Sao Carlos. Our results show that STBiodata took 24% less time (on average) for users to answer 10 queries compared to SpeciesLink in the scenarios described by the biodiversity specialists. We made an exploratory evaluation using a questionnaire to determine if STBioData helps the biodiversity specialist retrieve and link biodiversity and geospatial information.

There are a number of interesting avenues of future work. The first is improve the search process of STBioData including the option of searching the name of the species navigating by a taxonomy classification. Complementary with this, we are investigating a method of optimizing the quantity and quality of biodiversity data.

5.6 Remarks about the chapter

In this chapter, the research question 4, "How can we improve the location accuracy of biodiversity data?", was answered by an extension of the STBioData architecture to answer more advances geospatial queries. We formalized spatiotemporal characteristics of places (e.g. biomes and natural parks) to improve the data location of specimens from the SpeciesLink system.

A comparative analysis between STBioData and SpeciesLink showed that STBioData took 24% less time (on average) for users to answer ten queries. This mean that the possibility of creating queries that link information with other domains (e.g. geospatial and biodiversity) offers many alternatives to the domain experts in answering more advanced structured scientific questions in less time.

STBioData provided great benefit to biodiversity researchers that need geospatial results. The advantages of this integration include: allowing users to answer complex geospatial queries that could not be answered from the individual independent sources, and increasing speed in georeferencing and generating maps.

CONCLUSION

The main research question in this thesis was: "How can we integrate heterogeneous biodiversity semantic data using their spatiotemporal information and provenance?". Our hypothesis was that representing biodiversity data as Linked Data would improve the data integration from independent sources (if they share common ontology terms). In Chapter 4, the STBioData architecture was introduced. It automatically link spatiotemporal biodiversity data from heterogeneous data sources. It allows mapping between biodiversity data and the ontologies describing them and the generation of interactive geographic maps. A prototype, based on this architecture, was implemented, the STBioData system.

STBioData was tested using a representative datasets about biodiversity (46568 records) from the National Institute for Amazonian Research (INPA), the Brazilian Agricultural Research Corporation (EMBRAPA) and the IUCN Red List of Threatened SpeciesTM(IUCN). The STBioData prototypes were evaluated by biodiversity experts from the Laboratory of Molecular Biodiversity and Conservation of (UFScar) and Semantic Web specialists (USP). 90% of users were able to find information about the species they were after. Our testes showed that using the STBioData architecture was possible to integrate biodiversity data from different sources.

We also compared the task of finding relevant species data in the SpeciesLink site and in the STBioData system. Biologists conducted this experiment. Our results showed that the STBioData tool had better precision (24% improvement) and recall (22% improvement) when compared to the keyword based search, used by SpeciesLink.

We also investigated three other related research questions. First we investigated: "How can we improve the interoperability of biodiversity data?". The hypothesis was that representing biodiversity data, as Linked Data, would allow the use of more advanced and complex data queries, which were not possible before. In Chapter 2, we presented an architecture that uses a semantic search system to query biodiversity data in semantic web formats. We implemented a prototype and, with it, showed that the architecture produced results with better precision (28%),

recall (25%) and F1-measure (26%) than a standard biodiversity search tool (SpeciesLink), using real biodiversity data. We demonstrated that, once collection data is in RDF format, it is easy to integrate it with data from other independent data sources (also in RDF).

Second, we researched the question "How can provenance be modeled for multiple biodiversity use cases?". Our hypothesis was that adapting the W3C PROV standard, we could model provenance in the biodiversity domain and for multiple use cases. Indeed, as we showed in Chapter 3, this is possible. The W3C PROV specification defines a core data model for provenance. However, there is a lack of expressiveness using this generic W3C recommendation to model the different types of organisms that co-occur in time and space (geospatial relations). We presented a model for biodiversity data provenance (BioProv) based on the W3C PROV model. We defined a mapping example for biodiversity data from the IBT and generated RDF triples. Our results showed that the use of provenance information allowed experts to perform complex queries.

Finally, we asked "How can we improve the location accuracy of biodiversity data? ". Our hypothesis was that formalizing spatiotemporal characteristics, from biodiversity data, will allow more accurate data location. Indeed as we showed in Chapter 4 and Chapter 5 through our Linked Biodiversity Data architecture we added the possibility to enable easy searching, visualizing and downloading of accurate relevant data related to geospatial areas (biomes and natural parks) using GeoSPARQL functions.

The adoption of Semantic Web technologies not only improves biodiversity data representation but also the modeling of the knowledge contained in it. STBioData provided great benefits to biodiversity researches using it. The advantages of its use include answers for more complex, thus interesting, spatiotemporal queries that could not (or were hard to) be answered using traditional methods and increased speed in georeferencing data and generating maps.

6.1 Future work

There are a number of interesting avenues of future work that can emerge from our research, as described in the following:

First, this thesis has been mainly focused on linking biodiversity information from heterogeneous sources using their spatial location and temporal data. We reused ontologies and vocabularies to show the expressivity of the biodiversity data. We used the GeoSPARQL and Time ontologies for the spatiotemporal data. Furthermore, we aim to add the complete set of functions prescribed by GeoSPARQL to the search process. This could help to create more geospatial complex queries. Also, another idea is to add the sTSPARQL functions to allow for valid time support.

Second, the mapping and the query reformulator components should be optimized to

improve the interlinking between biodiversity vocabularies and ontologies with other domains. Interlinking is provided by RDF triples that establish a link between the entity identified by the subject with the entity identified by the object.

Third, there are different aspects of the proposed architecture which need to be assessed in future work such as (i) the performance in process complexity SPARQL and GeoSPARQL queries; (ii) the accuracy, precision and recall of the retrieved links in conjunction with other domains; and (iii) the usability and usefulness of STBioData, in the broadest sense of the word; studies can be conducted involving different types of stakeholders.

Fourth, we created a Provenance Model in the Biodiversity Domain (Bioprov). This model should be optimized in the cases when the biodiversity information is missing or incomplete. We can answer the question: How can we reconstruct and complete the provenance information in the biodiversity domain? This could help to improve the quality and quantity of the biodiversity data.

Fifth, we believe that the search process of the STBioData should be optimized including (i) the option to search for the name of the species navigating by a taxonomy classification. This could help to optimize the amount of time in the search process for the biodiversity specialists; (ii) the option to retrieve maps in different formats such as KML, GeoTiff, etc; (iii) the option to draw specific areas on the map with a few clicks; (iv) add other geospatial areas such as biomes, natural parks, localities, municipalities, etc. to the search process.

6.2 International collaborations

From August 2015 to June 2016, We have been working with the Data Science Laboratory of the University of Ghent, Belgium. This international collaboration helped us to accomplish the following activities:

- Identifying and analyzing of vocabularies and ontologies for linked geospatial data;
- Proposing a new model of provenance for biodiversity datasets (BioProv);
- Identifying of scientific use case: Molecular Identification of *Cladophora delicatula* Alga;
- Converting biodiversity data in the Semantic Web format (mapping). The mapping was implemented using the RML language and tested on a set of representative data about biodiversity and the model of provenance (BioProv).
- Performing experiments;
- Publishing two academic articles in international conferences:

- A Model of Provenance Applied to Biodiversity Datasets. 25th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2016). Paris, June 13-15, 2016 ([AMANQUI *et al.*, 2016b](#));
- Using Spatiotemporal Information to Integrate Heterogeneous Biodiversity Semantic Data. 16th International Conference on Web Engineering (ICWE2016) 6-9 June 2016, USI Lugano, Switzerland ([AMANQUI *et al.*, 2016a](#)).

Other extra activities included:

- Participating in research meetings;
- Attending the SuperMinds conference;
- Participating in RML Editor Evaluation;
- Participating in firefighting training and a Safety Steward UGent.

6.3 Publications

The research described in this thesis has lead to a total of 8 publications. Of those publications, I wrote 5 publications as first author, consisting of 2 journal articles and 3 international conferences. At the start of each chapter in this thesis, each article is listed.

Additionally, I have contributed as a co-author to 3 publications including 1 journal article and 2 conference publications.

- SWI: A Semantic Web Interactive Gazetteer to Support Linked Open Data. SD Cardoso, FK Amanqui, KJA Serique, JLC dos Santos, DA Moreira. *Future Generation Computer Systems* 54, 389-398 2015.
- A gazetteer for biodiversity data as a linked open data solution. SD Cardoso, KJ Serique, FK Amanqui, JLC Dos Santos, DA Moreira. *WETICE Conference (WETICE), 2014 IEEE 23rd International*, 435-440. 2014
- Social Participation Ontology:community documentation, enhancements and use examples. R Fabbri, RB de Luna, RAP Martins, FKM Amanqui, DA Moreira *arXiv preprint arXiv:1501.02662* 2015

BIBLIOGRAPHY

ABBAS, S.; OJO, A. Applying Design Patterns in URI Strategies – Naming in Linked Geospatial Data Infrastructure. In: **System Sciences (HICSS), 2014 47th Hawaii International Conference on**. [S.l.: s.n.], 2014. p. 2094–2103. Citation on page 57.

ALBUQUERQUE, A.; SANTOS, J. Campos dos; CASTRO, A. D. Ontobio: A Biodiversity Domain Ontology for Amazonian Biological Collected Objects. In: **System Sciences (HICSS), 2015 48th Hawaii International Conference on**. [S.l.: s.n.], 2015. p. 3770–3779. ISSN 1530-1605. Citation on page 44.

ALLEMANG, D.; HENDLER, J. **Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN 0123735564, 9780123735560. Citation on page 39.

AMANQUI, F.; VERBORGH, R.; MANNENS, E.; WALLE, R. Van de; MOREIRA, D. Using spatiotemporal information to integrate heterogeneous biodiversity semantic data. In: _____. **Web Engineering: 16th International Conference, ICWE 2016, Lugano, Switzerland, June 6-9, 2016. Proceedings**. Cham: Springer International Publishing, 2016. p. 525–530. ISBN 978-3-319-38791-8. Available: <http://dx.doi.org/10.1007/978-3-319-38791-8_41>. Citations on pages 36, 78, and 118.

AMANQUI, F. K.; NIES, T. D.; DIMOU, A.; VERBORGH, R.; MANNENS, E.; WALLE, R. V. D.; MOREIRA, D. A model of provenance applied to biodiversity datasets. In: **2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)**. [S.l.: s.n.], 2016. p. 235–240. Citations on pages 36, 74, 78, 81, 82, and 118.

AMANQUI, F. K.; POTENCIANO, G.; MOREIRA, D. A. **An Architecture to Create Geospatial Queries for the Biodiversity Domain, a Use Case about Endangered Species in Brazilian Biomes**. 2017. This article was submitted to the Information Systems Journal. Citation on page 36.

_____. **STBioData Linked Biodiversity Data using their Spatiotemporal Information, a Use Case about the Conservation Status of Amazon River Dolphins**. 2017. This article was submitted to the Information Systems Journal. Citations on pages 36, 100, and 113.

AMANQUI, F. K.; SERIQUE, K. J.; CARDOSO, S. D.; SANTOS, J. L.; ALBUQUERQUE, A.; MOREIRA, D. A. Improving biodiversity data retrieval through semantic search and ontologies. In: **Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on**. [S.l.: s.n.], 2014. v. 1, p. 274–281. Citations on pages 29, 30, 36, 81, 82, and 113.

ANDRADE, F. de; BAPTISTA, C. de S.; DAVIS CLODOVEUAUGUSTO, J. Improving geographic information retrieval in spatial data infrastructures. **GeoInformatica**, Springer US, v. 18, n. 4, p. 793–818, 2014. ISSN 1384-6175. Citation on page 63.

ANTONIOU, G.; ; ANTONIOU, G.; ANTONIOU, G.; HARMELEN, F. V.; HARMELEN, F. V. Web Ontology Language: OWL. In: **Handbook on Ontologies in Information Systems**. [S.l.]: Springer, 2003. p. 67–92. Citation on page 39.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. **Modern Information Retrieval**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999. ISBN 020139829X. Citation on page 42.

BASKAUF, S. J.; WEBB, C. O. Darwin-SW: Darwin Core-based Terms for Expressing Biodiversity Data as RDF. **Semantic Web**, IOS Press, v. 7, n. 6, p. 629–643, 2016. Citations on pages 45, 78, and 81.

BERKLEY, C.; BOWERS, S.; JONES, M.; MADIN, J.; SCHILDHAUER, M. Improving data discovery for metadata repositories through semantic search. In: . [S.l.: s.n.], 2009. p. 1152–1159. Citation on page 42.

BERNERS-LEE, T. Linked Data - Design Issues. **W3C**, ACM Press, n. 09/20, 2006. Available: <http://www.w3.org/DesignIssues/LinkedData.html>. Citations on pages 33 and 100.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, v. 284, n. 5, p. 34–43, May 2001. Available: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Citations on pages 33, 35, and 40.

BOLEY, H.; TABET, S.; WAGNER, G. Design rationale of ruleml: A markup language for semantic web rules. In: . [S.l.: s.n.], 2001. p. 381–401. Citations on pages 35 and 41.

BUTTIGIEG, P.; MORRISON, N.; SMITH, B.; MUNGALL, C.; LEWIS, S.; CONSORTIUM the E. The environment ontology: contextualising biological and biomedical entities. **Journal of Biomedical Semantics**, v. 4, n. 1, p. 43+, 2013. ISSN 2041-1480. Available: <http://dx.doi.org/10.1186/2041-1480-4-43>. Citation on page 44.

CARDOSO, S. D.; AMANQUI, F. K.; SERIQUE, K. J.; SANTOS, J. L. dos; MOREIRA, D. A. Swi: A semantic web interactive gazetteer to support linked open data. **Future Generation Computer Systems**, n. 0, p. –, 2015. ISSN 0167-739X. Available: <http://www.sciencedirect.com/science/article/pii/S0167739X15001818>. Citations on pages 29 and 77.

CELINO, I. Human computation vgi provenance: Semantic web-based representation and publishing. **Geoscience and Remote Sensing, IEEE Transactions on**, v. 51, n. 11, p. 5137–5144, Nov 2013. ISSN 0196-2892. Citations on pages 59, 62, and 74.

COMMUNITY, G. Vocabulary Management Group VoMaG. In: . [s.n.], 2017. Available: <http://community.gbif.org/pg/groups/21382/vocabulary-management/>. Citation on page 73.

CONSORTIUM, O. G. Geosparql - A Geographic Query Language for RDF Data. In: . [s.n.], 2017. Available: <http://www.opengeospatial.org/standards/geosparql>. Citations on pages 58, 77, 78, and 81.

CYGANIAK, R.; JENTZSCH, A. **The Linking Open Data cloud diagram**. 2014. <http://lod-cloud.net/>. Available: <http://lod-cloud.net/>. Citations on pages 33 and 34.

DEVEDZIC, V. Education and the semantic web. **Int. J. Artif. Intell. Ed.**, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 14, n. 2, p. 165–191, Apr. 2004. ISSN 1560-4292. Available: <http://dl.acm.org/citation.cfm?id=1434858.1434860>. Citations on pages 35 and 39.

DEVELOPERS, G. The Google Geocoding API - Google Maps API Web Services — Google Developers. 2017. Available: <<https://developers.google.com/maps/documentation/geocoding/>>. Citations on pages 86 and 106.

DIMOU, A.; SANDE, M. V.; COLPAERT, P.; VERBORGH, R.; MANNENS, E.; WALLE, R. Van de. Rml:a generic language for integrated RDF mappings of heterogeneous data. In: **Proceedings of the 7th Workshop on Linked Data on the Web**. [S.l.: s.n.], 2014. Citations on pages 63 and 66.

Eclipse RDF4J. **The RDF4J Server REST API**. 2017. <<http://docs.rdf4j.org/rest-api/>>. Accessed: 2017-09-19. Citations on pages 87 and 105.

EMBRAPA.BR. **The Brazilian Agricultural Research Corporation - EMBRAPA**. <<https://www.embrapa.br/en>>. Accessed: 2017-09-30. Citations on pages 29 and 76.

GAIGNARD, A.; MONTAGNAT, J.; GIBAUD, B.; FORESTIER, G.; GLATARD, T. Domain-specific summarization of life-science e-experiments from provenance traces. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 29, p. 19 – 30, 2014. ISSN 1570-8268. Life Science and e-Science. Available: <<http://www.sciencedirect.com/science/article/pii/S1570826814000493>>. Citation on page 78.

GBIF.ORG. The Global Biodiversity Information Facility GBIF. In: . [S.l.: s.n.]. Accessed: 2017-09-30. Citations on pages 29, 42, 60, and 73.

GOELDI.BR museu. **The Emilio Gueldi Museum in Para MPEG**. <<http://www.museu-goeldi.br/portal/>>. Accessed: 2017-09-30. Citation on page 29.

GREEN, J.; HASTINGS, A.; ARZBERGER, P.; AYALA, F.; COTTINGHAM, K.; CUDDINGTON, K.; DAVIS, F.; DUNNE, J.; FORTIN, M.; GERBER, L.; NEUBERT, M. Complexity in ecology and conservation: Mathematical, statistical, and computational challenges. **Bioscience**, v. 55, p. 501–510, 2005. Citation on page 33.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. **Int. J. Hum.-Comput. Stud.**, Academic Press, Inc., Duluth, MN, USA, v. 43, n. 5-6, p. 907–928, Dec. 1995. ISSN 1071-5819. Available: <<http://dx.doi.org/10.1006/ijhc.1995.1081>>. Citation on page 35.

GÜTING, R. H. Graphdb: Modeling and querying graphs in databases. In: **Proceedings of the 20th International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 297–308. ISBN 1-55860-153-8. Available: <<http://dl.acm.org/citation.cfm?id=645920.672980>>. Citation on page 46.

HASNAIN, A.; MEHMOOD, Q.; ZAINAB, S. Sana e; SALEEM, M.; WARREN, C.; ZEHR, D.; DECKER, S.; REBHOLZ-SCHUHMANN, D. Biofed: federated query processing over life sciences linked open data. **Journal of Biomedical Semantics**, v. 8, n. 1, p. 13, 2017. ISSN 2041-1480. Available: <<http://dx.doi.org/10.1186/s13326-017-0118-0>>. Citation on page 77.

HEATH, T.; BIZER, C. **Linked Data: Evolving the Web into a Global Data Space**. 1st. ed. [S.l.]: Morgan & Claypool, 2011. ISBN 9781608454303. Citation on page 34.

INC., O. G. C. **OpenGIS Implementation Standard for Geographic information - Simple feature access**. 1.2.1. ed. [S.l.: s.n.], 2011. Citation on page 77.

INPA, N. I. for A. R. Specieslink System. In: . [s.n.], 2017. Available: <<http://splink.cria.org.br/>>. Citations on pages 42 and 76.

INPA.BR. **The National Institute for Amazonian Research INPA**. <<http://portal.inpa.gov.br/>>. Accessed: 2017-09-30. Citations on pages 29, 60, and 76.

INPE.BR. **The National Institute for Space Research INPE**. <<http://www.inpe.br/>>. Accessed: 2017-09-30. Citation on page 29.

KAUPPINEN, T.; ESPINDOLA, G. M. de. Linked Open Science—communicating, sharing and evaluating data, methods and results for executable papers. **Proceedings of the International Conference on Computational Science (ICCS 2011)**, *Procedia Computer Science*, v. 4, n. 0, p. 726–731, 2011. ISSN 1877-0509. Citations on pages 44 and 52.

KAUPPINEN, T.; ESPINDOLA, G. M. de; JONES, J.; SANCHEZ, A.; GRÄLER, B.; BARTOSCHEK, T. Linked Brazilian Amazon Rainforest Data. **Semantic Web Journal**, IOS Press, v. 5, n. 2, 2014. Citation on page 74.

KOUBARAKIS, M.; KARPATHIOTAKIS, M.; KYZIRAKOS, K.; NIKOLAOU, C.; SIOUTIS, M. Data Models and Query Languages for Linked Geospatial Data. In: EITER, T.; KREN-NWALLNER, T. (Ed.). **Reasoning Web. Semantic Technologies for Advanced Query Answering**. [S.l.]: Springer Berlin / Heidelberg, 2012, (Lecture Notes in Computer Science, v. 7487). p. 290–328. ISBN 978-3-642-33157-2. Citation on page 59.

KUHN, W.; KAUPPINEN, T.; JANOWICZ, K. Linked data - a paradigm shift for geographic information science. In: **Geographic Information Science**. Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8728). p. 173–186. ISBN 978-3-319-11592-4. Available: <http://dx.doi.org/10.1007/978-3-319-11593-1_12>. Citations on pages 33 and 61.

LACASTA, J.; LOPEZ-PELLICER, F.; RENTERIA-AGUALIMPIA, W.; NOGUERAS-ISO, J. Improving the visibility of geospatial data on the web. In: **Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on**. [S.l.: s.n.], 2014. p. 155–164. Citation on page 113.

LASSILA, O.; SWICK, R. R.; WIDE, W.; CONSORTIUM, W. **Resource Description Framework (RDF) Model and Syntax Specification**. 1998. Citations on pages 37, 63, 73, and 99.

LBA.BR. **The Large-Scale Biosphere-Atmosphere Experiment in Amazonia LBA**. <<http://lba.inpa.gov.br>>. Accessed: 2017-09-30. Citation on page 60.

LEHMANN, J.; ISELE, R.; JAKOB, M.; JENTZSCH, A.; KONTOKOSTAS, D.; MENDES, P. N.; HELLMANN, S.; MORSEY, M.; KLEEF, P. van; AUER, S.; BIZER, C. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. **Semantic Web Journal**, v. 6, n. 2, p. 167–195, 2015. Available: <http://jens-lehmann.org/files/2014/swj_dbpedia.pdf>. Citations on pages 45, 78, and 81.

LEON, A. de; WISNIEWKI, F.; VILLAZÓN-TERRAZAS, B.; CORCHO, O. Map4rdf - faceted browser for geospatial datasets. In: **PMOD workshop**. [s.n.], 2012. Oeg. Available: <<http://oa.upm.es/14511/>>. Citation on page 77.

MAGNUSSON, W. E. Biodiversity and integrated environmental monitoring. In: **Program for Planned Biodiversity and Ecosystem Research (PPBio)**. [S.l.: s.n.], 2013. v. 1, p. 273–289. Citations on pages 29, 60, 62, and 63.

MALAVERRI, J. E. G.; MEDEIROS, C. B.; LAMPARELLI, R. C. A Provenance Approach to Assess the Quality of Geospatial Data. In: **Proceedings of the 27th Annual ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2012. (SAC '12), p. 2043–2044. ISBN 978-1-4503-0857-1. Citation on page 62.

MANGOLD, C. A survey and classification of semantic search approaches. **Int. J. Metadata Semant. Ontologies**, Inderscience Publishers, Inderscience Publishers, Geneva, SWITZERLAND, v. 2, n. 1, p. 23–34, Sep. 2007. ISSN 1744-2621. Available: <<http://dx.doi.org/10.1504/IJMSO.2007.015073>>. Citation on page 41.

MCTIC, B. M. of S. **The Reference Center on Environmental Information - CRIA**. 2017. <<http://www.cria.org.br/>>. Accessed: 2017-07-30. Citation on page 60.

MIKOWSKI, M.; POWELL, J. **Single Page Web Applications: JavaScript End-to-end**. 1st. ed. Greenwich, CT, USA: Manning Publications Co., 2013. ISBN 1617290750, 9781617290756. Citation on page 86.

MIRON, A.; GENSEL, J.; VILLANOVA-OLIVER, M. Discovery of spatial links between individuals on the geospatial semantic web. In: **Geoinformatics, 2009 17th International Conference on**. [S.l.: s.n.], 2009. p. 1–6. Citation on page 41.

MOREAU, L.; FREIRE, J.; FUTRELLE, J.; MCGRATH, R. E.; MYERS, J.; PAULSON, P. The open provenance model: An overview. In: **Provenance and Annotation of Data and Processes**. [S.l.]: Springer, 2008. p. 323–326. Citation on page 62.

MOREAU, L.; HUYNH, T. D.; MICHAELIDES, D. An online validator for provenance: Algorithmic design, testing, and api. In: **Proceedings of the 17th International Conference on Fundamental Approaches to Software Engineering - Volume 8411**. New York, NY, USA: Springer-Verlag New York, Inc., 2014. p. 291–305. ISBN 978-3-642-54803-1. Available: <http://dx.doi.org/10.1007/978-3-642-54804-8_20>. Citation on page 83.

MOREAU, L.; Missier (Eds.), P.; W3C Provenance Working Group. **PROV-DM: The PROV Data Model**. W3C. 2013. Citations on pages 60 and 66.

_____. **PROV: PROV Model Primer**. W3C. 2013. Citations on pages 59, 63, 64, 81, and 82.

MOURA, A. M. de C.; PORTO, F.; VIDAL, V.; MAGALHÃES, R. P.; MAIA, M.; POLTOSI, M.; PALAZZI, D. A semantic integration approach to publish and retrieve ecological data. **International Journal of Web Information Systems**, v. 11, n. 1, p. 87–119, 2015. Available: <<http://dx.doi.org/10.1108/IJWIS-08-2014-0028>>. Citation on page 77.

NIKOLAOU, C.; DOGANI, K.; BERETA, K.; GARBIS, G.; KARPATHIOTAKIS, M.; KYZIRAKOS, K.; KOUBARAKIS, M. Sextant: Visualizing time-evolving linked geospatial data. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 35, Part 1, p. 35 – 52, 2015. ISSN 1570-8268. Geospatial Semantics. Available: <<http://www.sciencedirect.com/science/article/pii/S1570826815000864>>. Citation on page 77.

NYGB. **The New York Botanical Garden (NYGB)**. 2017. <<http://www.nybg.org>>. Accessed: 14-September-2017. Citation on page 60.

OGC. Well-known text representation of coordinate reference systems. In: . [s.n.], 2017. Available: <<http://www.opengespatial.org/standards/wkt-crs>>. Citation on page 58.

PRUD'HOMMEAUX, E.; SEABORNE, A. **SPARQL Query Language for RDF**. [S.l.], 2006. Available: <http://www.w3.org/TR/rdf-sparql-query/>. Citations on pages 38, 77, and 100.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2008. ISBN 3-900051-07-0. Available: <http://www.R-project.org>. Citation on page 74.

R, M. M.; CALVANESE, D. **Quest, an OWL 2 QL Reasoner for Ontology-Based Data Access**. [S.l.]: KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, Bolzano, Italy, 2012. Citation on page 39.

SAHOO, A. S. S. **Provenir Ontology: Towards a Framework for eScience Provenance Management**. 2009. 15-17 p. Citation on page 63.

SANTOS, J. L. C. dos. **A Biodiversity Information System in an Open Data/Metadatabase Architecture**. Phd Thesis (PhD Thesis) — University of Twente, Enschede, 2003. Citations on pages 29, 40, 60, and 74.

SANTOS, V. D.; BAIAO, F. A.; TANAKA, A. An architecture to support information sources discovery through semantic search. In: **Information Reuse and Integration**. [S.l.: s.n.], 2011. Citation on page 42.

SHARMA, A. K.; DUHAN, N.; SHARMA, B. A semantic search system using query definitions. In: **Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia**. New York, NY, USA: ACM, 2010. (IITM '10), p. 279–283. ISBN 978-1-4503-0408-5. Available: <http://doi.acm.org/10.1145/1963564.1963613>. Citation on page 42.

SIERRA, S.; VANDERHART, L. **ClojureScript: Up and Running Functional Programming for the Web**. 2nd. ed. [S.l.]: O'Reilly Media, Inc., 2016. ISBN 1491924993, 9781491924990. Citations on pages 86 and 105.

SKEVAKIS, G.; MAKRIS, K.; KALOKYRI, V.; ARAPI, P.; CHRISTODOULAKIS, S. Metadata management, interoperability and linked data publishing support for natural history museums. **International Journal on Digital Libraries**, v. 14, n. 3, p. 127–140, 2014. ISSN 1432-1300. Available: <http://dx.doi.org/10.1007/s00799-014-0114-2>. Citation on page 77.

SOUSA, R. B.; CUGLER, D. C.; MALAVERRI, J. G.; MEDEIROS, C. B. A provenance-based approach to manage long term preservation of scientific data. In: **Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on**. [S.l.: s.n.], 2014. p. 162–133. Citation on page 62.

SPECIES, T. I. R. L. of T. IUCN. 2017. <http://www.iucnredlist.org/>. Online; accessed 14-September-2017. Citations on pages 75 and 76.

STADLER, C.; MARTIN, M.; AUER, S. Exploring the web of spatial data with facete. In: **Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion**. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014. (WWW Companion '14), p. 175–178. ISBN 978-1-4503-2745-9. Citations on pages 77 and 113.

TAXIDOU, I.; NIES, T. D.; VERBORGH, R.; FISCHER, P.; MANNENS, E.; WALLE, R. Van de. Modeling Information Diffusion in Social Media as Provenance with W3C PROV. In: **Proceedings of the 6th International Workshop on Modeling Social Media**. [S.l.: s.n.], 2015. Citation on page 60.

W3C. **OWL 2 Web Ontology Language Profiles (Second Edition)**. [S.l.]: W3C Recommendation, 2012. Available at <http://www.w3.org/TR/owl2-profiles/#OWL_2_RL>. Citation on page 39.

_____. The PROV Ontology (PROV-O). In: . [s.n.], 2013. Available: <<https://www.w3.org/TR/prov-o/>>. Citation on page 60.

_____. Time Ontology in OWL. In: . [s.n.], 2017. Available: <<https://www.w3.org/TR/owl-time/>>. Citation on page 82.

WANG, S.; PADMANABHAN, A.; MYERS, J. D.; TANG, W.; LIU, Y. Towards provenance-aware geographic information systems. In: **Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**. New York, NY, USA: ACM, 2008. (GIS '08), p. 70:1–70:4. ISBN 978-1-60558-323-5. Available: <<http://doi.acm.org/10.1145/1463434.1463515>>. Citation on page 62.

WARMERDAM, F. The geospatial data abstraction library. In: _____. **Open Source Approaches in Spatial Data Handling**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 87–104. ISBN 978-3-540-74831-1. Available: <https://doi.org/10.1007/978-3-540-74831-1_5>. Citation on page 82.

WEIBEL, S. L.; KOCH, T. **Dublin Core Metadata Initiative (DCMI)**. 2000. Citation on page 62.

WICK, M. **Geonames Ontology**. 2015. Available: <<http://www.geonames.org/about.html>>. Citation on page 45.

WIEGAND, N.; GROVE, R.; WILSON, J.; KOLAS, D. Querying geospatial data over the web: A geosparql interface. In: **Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data**. New York, NY, USA: ACM, 2007. (GeoRich'14), p. 4:1–4:6. ISBN 978-1-4503-2978-1. Available: <<http://doi.acm.org/10.1145/2619112.2619116>>. Citation on page 58.

XIONG, J.; HUANG, W.; JIN, C. An ontology-based semantic search approach for geosciences. In: . [S.l.: s.n.], 2009. v. 3, p. 87–90. Citation on page 42.

YUAN, J.; YUE, P.; GONG, J.; ZHANG, M. A linked data approach for geospatial data provenance. **Geoscience and Remote Sensing, IEEE Transactions on**, v. 51, n. 11, p. 5105–5112, Nov 2013. ISSN 0196-2892. Citation on page 62.

ZAVERI, A.; RULA, A.; MAURINO, A.; PIETROBON, R.; LEHMANN, J.; AUER, S. Quality assessment for linked data: A survey. **Semantic Web**, IOS Press, v. 7, n. 1, p. 63–93, 2016. Citation on page 77.

ZHANG, Y.; LI, C.; LIU, S.; WEN, F.; DU, L.; HE, H. A unified approach to automate geospatial data retrieval using semantic web technologies. In: **2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)**. [S.l.: s.n.], 2016. p. 1–6. Citation on page 113.

ZHAO, C.; WANG, J.; HU, W.; YU, X.; WANG, X. An ontology-based semantic search model study. In: . [S.l.: s.n.], 2010. p. 182 –185. Citations on pages [42](#), [44](#), and [62](#).

ZIEGLER, P.; DITTRICH, K. R. **Data Integration – Problems, Approaches, and Perspectives**. 2007. Citation on page [33](#).