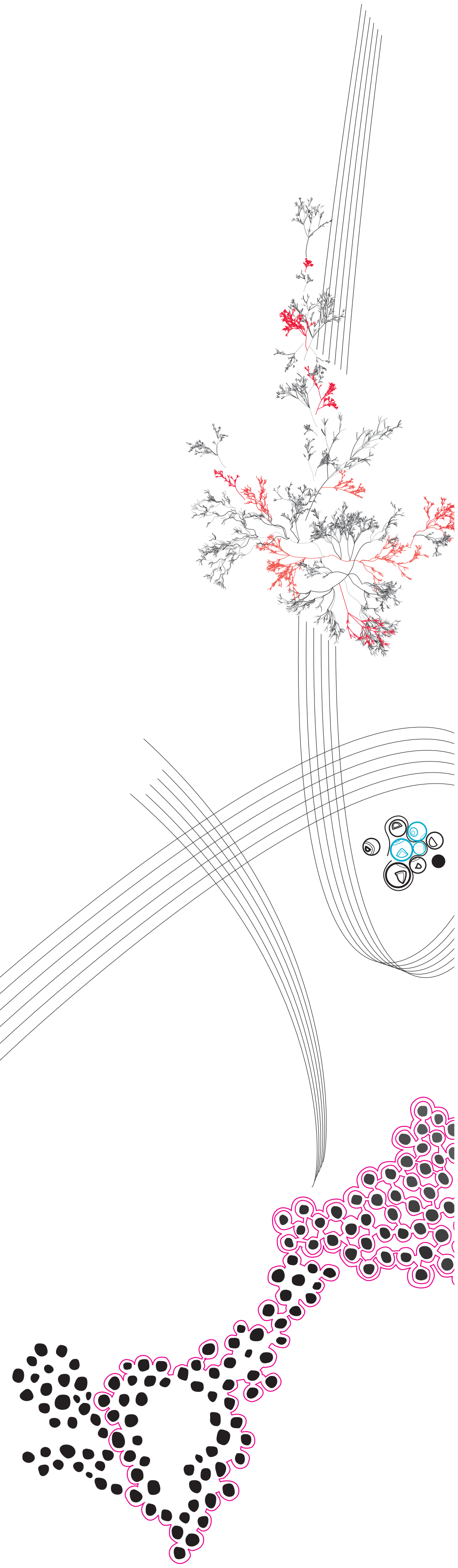
MSc Computer Science
Final Project



Unlocking the Butterfly Effect: Finding Hidden Connections in Barcant Butterfly Collection with Ontology Matching and Knowledge Graphs

Rakshitha VijayKumar

Supervisor: Dr. Faiza A. Bukhsh
Dr. Shenghui Wang,
Dr. Andreas Weber

November, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

Biodiversity data is essential for conservation and ecosystem management, yet its rapid accumulation from diverse sources presents challenges for integration and analysis. This study addresses these challenges by applying ontology matching and knowledge graph techniques to the Barcant Butterfly Collection (BBC), one of the largest butterfly archives in the Caribbean. The research focuses on transforming and standardizing the BBC dataset using the Darwin Core (DwC) framework to facilitate interoperability with global biodiversity information systems.

Through the construction of a domain-specific ontology for the BBC, aligned with existing biodiversity ontologies such as the Biological Collections Ontology (BCO) and Taxonomic Rank Ontology (TaxRank), this work creates a structured, interoperable knowledge base. The ontology alignment process enhances the consistency and integration of taxonomic, ecological, and geographic data. A knowledge graph is generated to reveal hidden connections and patterns within the collection, enabling advanced querying and analysis.

The results demonstrate the potential of these methodologies to unlock valuable insights from legacy biological collections, enhancing their utility for conservation efforts and biodiversity research. By aligning traditional taxonomic data with modern computational techniques, this research contributes to the broader field of biodiversity informatics, setting a precedent for the digitization and semantic enrichment of biological collections worldwide.

Keywords: Barcant Butterfly Collection, Ontology alignment, knowledge graphs, biodiversity informatics.

Chapter 1

Introduction

Biodiversity data is crucial for conservation and ecosystem management. It helps identify endangered species and at-risk environments, providing key insights for creating effective conservation strategies tailored to specific ecosystem's needs. Advances in information technology and the expansion of open-access data-sharing platforms have significantly increased the availability of biodiversity data. In recent years, the volume of data has also surged, largely due to crowd sourcing initiatives such as citizen science programs, environmental monitoring networks, and contributions from research institutions. These various sources provide detailed information on genetic diversity, species distributions, taxonomic lists, and how species interact within ecosystems. For example, citizen science programs often add real-time data about local ecosystems, offering detailed observations that complement large-scale research. The integration of these efforts ensures that conservation strategies are built on the most comprehensive data available, thereby enabling ecosystem sustainability [1].

However, the rapid accumulation of data through such heterogeneous techniques often leads to discrepancies and ambiguities within the records, so it becomes important to establish standards for data integration to ensure semantic interoperability. A unified approach allows for the seamless combination of data, providing valuable insights into the evolutionary history of life on Earth and deepening our understanding of species-specific threats, habitat changes, and biodiversity hotspots [2]. These insights enable the development of proactive strategies that protect vulnerable species and their ecosystems. Furthermore, accurate data on species distributions and ecosystems is crucial for predicting the impacts of climate change on biodiversity, allowing for the creation of strategies that safeguard habitats [3] [4].

The global digitization of biodiversity collections, led by museums, has significantly transformed research methodologies. Traditionally, museums focused on cataloging physical specimens, but now they are increasingly digitizing these collections using advanced methods like DNA sequencing, high-resolution imaging, and data annotation. This digital shift, which combines classic taxonomy with modern genomics, has introduced both new opportunities and challenges in managing and analyzing the massive datasets that result from this process [5]. One of the major challenges is the localized or narrow focus of many organizations when compiling and depicting data [6]. Each institution often uses its own framework, leading to inconsistencies across datasets. For example, identical species observations may be attributed to multiple sources, or the same scientific name may be applied to different species, leading to confusion and duplication in biodiversity records. Moreover, the sheer abundance of disparate data sources, ranging from museum collections to research institutions, exacerbates this issue. Integrating these heterogeneous datasets presents significant challenges due to variations in observational scales, data collection methodologies, and terminologies. Harmonizing this data collected from such various streams requires the development of universal standards for data integration and classification to ensure consistency and accuracy.

The **Barcant Butterfly Collection (BBC)**, used in this research, is one such highly valuable dataset in need of standardization. It documents a wide range of butterfly species from Trinidad and Tobago. It is one of the most comprehensive butterfly archives in the Caribbean, containing many endemic species. Assembled by Malcolm Barcant over several decades, the collection not only showcases the rich diversity of butterfly species but also provides critical insights into the ecological dynamics of the Caribbean [7]. Despite its historical and scientific significance, the full potential of the Barcant Butterfly Collection remains underutilized in the context of big data and advanced computational techniques. For this dataset to reach its maximum utility, it must be integrated into global biodiversity information systems that utilize standardized frameworks, enabling seamless data interoperability and more sophisticated analyses. Such integration is essential for facilitating accurate comparisons, drawing reliable conclusions, and enabling meta-studies that combine datasets across researchers, regions, and time periods.

The integration of traditional taxonomic knowledge with modern data science methods, such as ontology alignment and knowledge graph construction, can greatly enhance the usability of the Barcant Butterfly Collection. Ontology alignment is particularly crucial for harmonizing diverse taxonomic frameworks. By aligning these frameworks, we can ensure consistent classification across various data sources, making it easier to integrate and compare information [8]. By employing these techniques, we can bridge the gap between analog historical records and contemporary digital databases. To accurately capture the nuances of butterfly biodiversity within the collection and tackle the challenges of data integration, the research will be guided by the following research questions:

> **RQ1**: *Concepts and Relationships: What specific ontological concepts and relationships are crucial for developing a comprehensive Barcant Butterfly Collection ontology that effectively captures the nuances of butterfly biodiversity?*
>
> **RQ1.a**: *Selection of standard vocabulary and terminologies from existing ontologies: Which standardized sets of terms can be efficiently repurposed or adopted from existing ontologies? And, how can these domain-specific lexicons be leveraged to facilitate the ontology matching processes without necessitating explicit redefinition or redundancies?*
>
> **RQ1.b**: *Identification of an ideal biodiversity ontology for alignment: Among the wide range of biodiversity ontologies available, which specific ontology aligns most seamlessly with the intricacies of the Barcant Butterfly Collection, ensuring optimal relevance and applicability?*

> **RQ2**: *Evaluation of the Barcant Butterfly Collection ontology: How can the Barcant Butterfly Collection ontology be effectively evaluated, and which specific queries are suitable for assessing its performance, efficacy, and practical applicability?*

The application of advanced techniques to the Barcant Butterfly Collection has the potential to serve as a model for the digitization and analysis of biological collections worldwide. Numerous museums and research institutions house extensive archives of specimens, each with distinct historical and scientific value. However, much like the Barcant Collection, many of these collections remain underutilized, with their data trapped in records that are not easily accessible or analyzable [9]. By demonstrating the value of integrating modern data science methodologies with traditional biological archives, this research aims to inspire similar digitization efforts across other institutions, contributing to a broader, global understanding of biodiversity. This aligns with the "Butterfly Effect" metaphor, wherein small contributions to data accessibility can significantly enhance the understanding and analysis of regional specimens, potentially leading to far-reaching implications. As global biodiversity faces increasing threats from climate change and habitat destruction, the need for novel conservation strategies is more important than ever. By leveraging ontology and knowledge graphs, this research unlocks the hidden potential of the Barcant Butterfly Collection, transforming it into a valuable resource for global biodiversity conservation.

The remainder of this thesis is organized as follows: Chapter 2 provides an overview of fundamental concepts related to ontology and knowledge graphs. Chapter 3 presents a comprehensive review of existing literature and previous work in the field, highlighting their relevance and contributions to this research. Chapter 4 talks about the dataset used in the research. Chapter 5 outlines the methodology, detailing the implementation process for addressing the research questions. chapter 6 demonstrates the practical use of the developed methodology in real-world scenarios, showcasing its effectiveness and relevance. Finally, Chapter 7 concludes the study and suggests potential directions for future research.

Chapter 2

Background

2.1 Ontology

In information science and computing, ontology represents a formal framework that defines a set of concepts and their relationships within a specific domain [10]. This structured approach facilitates data integration, interoperability, and reasoning, making it easier for systems to work with data coherently and efficiently. In computational fields, ontologies are used to provide a shared and common understanding of a domain. This common understanding allows for consistent communication between people and software agents, ensuring that different systems interpret data in the same way despite variations in structure or context. Ontologies are crucial in areas such as artificial intelligence, semantic web technologies, and bioinformatics, where they enable semantic interoperability and enhance data management. They are designed as comprehensive schemas that categorize and define the relationships and properties of various concepts within a domain. Ontologies serve as a detailed "specification of a conceptualization" that integrates information from both structured and unstructured sources [11]. This approach allows ontologies to encapsulate domain-specific semantics, providing a richer understanding of how concepts relate to one another.

A significant advantage of ontologies is their ability to define a wide array of relationships and properties, which enhances their applicability in complex scenarios such as reasoning and inference. Ontologies can define classes, relationships between these classes, and attributes, facilitating advanced querying and data integration [12]. Ontologies provide descriptions of the following elements [13, 14]:

- Classes or “Things” within different domains of interest
- Relationships among said “Things”
- Properties or attributes the “Things” should possess

Biodiversity Ontology is a specialized application of ontologies in biodiversity science that plays a critical role in organizing and standardizing the vast and complex datasets associated with species, ecosystems, and genetic information. These ontologies define clear relationships between organisms, their habitats, ecological interactions, and environmental factors, creating a structured framework that enhances data interoperability [16]. For instance, the Environment Ontology (ENVO) provides a standardized vocabulary for describing environmental features, ecosystems, and

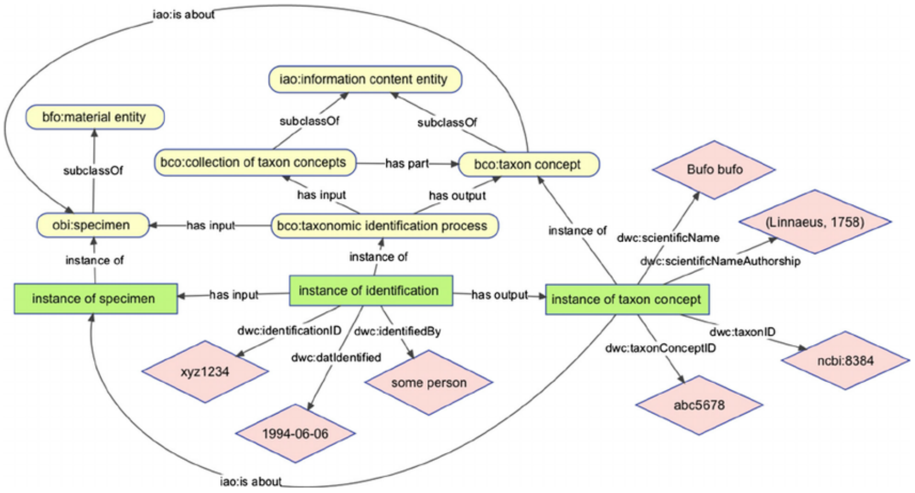


Figure 2.1: Ontological representation of the identification process in the Biological Collections Ontology. Ontologies are abbreviated as "bfo" for Basic Formal Ontology, "bco" for Biological Collections Ontology, and "dwc" for Darwin Core [15]

habitat characteristics. Biodiversity ontologies play a crucial role in the classification and naming of new species, ensuring standardized descriptions that can be shared across numerous scientific disciplines [17]. They also facilitate modeling of complex ecological processes, such as species interactions and migration patterns. The Open Biomedical Ontologies (OBO) Foundry, which includes biodiversity-related ontologies, exemplifies how these tools can integrate data across life sciences.

The use of ontologies enhances the management and analysis of large-scale datasets, such as those collected by the Global Biodiversity Information Facility (GBIF)[1]. Recent advancements in machine learning and artificial intelligence have further amplified the impact of ontologies [18]. These technologies leverage ontologies for automated species identification, ecosystem monitoring, and predictive modeling, offering new possibilities for biodiversity conservation and management in response to global challenges. Ontologies provide a sophisticated approach to knowledge representation, serving as detailed schemas that categorize and interconnect concepts.

2.2 Ontology Matching

Ontology matching is the process of identifying correspondences between semantically related entities in different ontologies. These correspondences may represent relationships such as equivalence, consequence, or disjointness among ontology entities. Ontology entities typically refer to the named elements within an ontology, such as classes, properties, or individuals. However, these entities can also encompass more complex structures like formulas, concept definitions, or term-building expressions [19].

Matchers are the core components of the ontology matching process, responsible for generating correspondences based on various factors such as entity labels and

[1] https://www.gbif.org/

structural relationships [20]. Different types of matchers include:

- **Basic matchers** - Focus on point-to-point mappings using lexical or structural similarities. For example, matching "Mammal" in one ontology with "Mammalia" in another.
- **Terminological matchers** - Explore concept labels through string-based or linguistic methods. For instance, matching "Danaus plexippus" (Latin name) with "Monarch Butterfly" (common name) across biodiversity ontologies.
- **Structural matchers,** - Analyze relationships between concepts within ontologies. For example, recognizing that "Monarch Butterfly" is a subclass of "Insect" and relating these structural relationships across different biodiversity classifications.

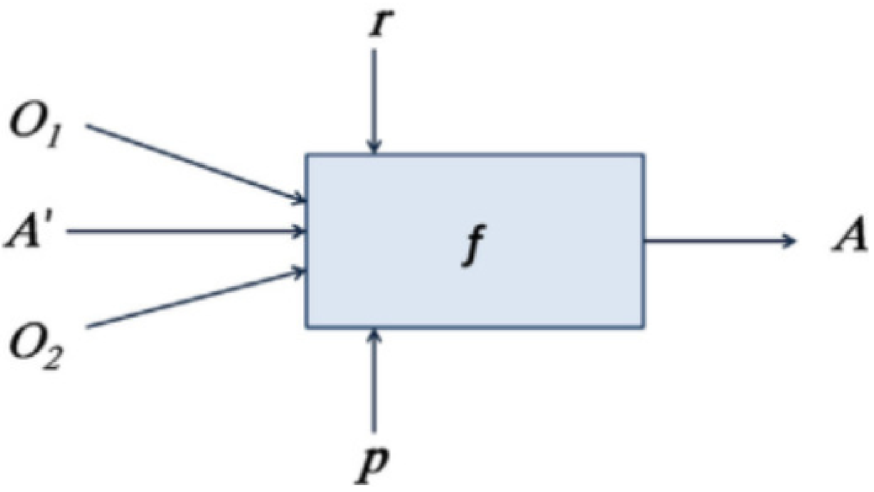


Figure 2.2: Flowchart of Ontology Matching Process

The flowchart in fig. 2.2 illustrates the ontology matching process. In this workflow, two ontologies, labeled O_1 and O_2, are subjected to a matching procedure governed by a function f, which generates the final matching result M. The matching function is influenced by several inputs, the ontologies O_1 and O_2, previously matched entities or intermediate results, denoted as A'. Additionally, the process is directed by a set of guidelines or relationships, represented as r and p , which may consist of logical rules or similarity metrics essential for the matching task. These parameters may include thresholds, weights, or other preferences that define how the function operates. The function f synthesizes these inputs to produce the final matched ontology A, ensuring semantic alignment between O_1 and O_2.

Combining different matcher types improves mapping accuracy by addressing both terminological and structural variations [21]. For instance, when aligning two ontologies—O1 and O2—matchers identify relationships such as one-to-one (1:1) equivalences (e.g., ’Species’ and ’Specimen’), one-to-many relationships (e.g., ’Mammal’ as a subclass of both ’Vertebrate’ and ’Warm-blooded Animal’), or many-to-many relationships where partial overlap exists between entities (e.g., "Animal" and "Organism"). Recent advances in ontology matching emphasize the importance of integrating various matching techniques to improve both accuracy and scalability [22]. In bioinformatics and other complex fields, error detection during matching is crucial for ensuring reliable, high-quality results.

As diverse ontologies continue to be developed without established mappings between them, effective ontology matching techniques become increasingly important. One innovative approach is presented by Xingsi Xue et al. [23], who utilized an Evolutionary Algorithm (EA) for ontology alignment. While traditional EA has been effective, it often struggles with large-scale ontologies and can get trapped in local optima. To address these limitations, the authors introduced the Adaptive Compact EA (ACEA), which uses semantic reasoning to filter out negative correspondences, thus reducing the search space. ACEA dynamically adjusts the search direction to explore previously unexplored regions, improving the overall effectiveness of the alignment. The use of multiple Probability Matrices (PMs) helps guide the search process. The study's findings indicate that ACEA-based techniques outperform other EA-based methods, although they may still have limitations, particularly in detecting correspondences.

Karam et al. [24] explored the matching of various biodiversity ontologies, noting that many of these ontologies lack inherent connections. One notable exception is a manual mapping between the Environment Ontology (ENVO) and a portion of the Semantic Web for Earth and Environment Technology Ontology (SWEET) subdomain. The reference alignments in this study were created using consensus mappings from existing systems, manually validated mappings, and expert-generated mappings. The results revealed that while most systems handled consensus mappings well, expert mappings posed greater challenges. This highlights the need for specialized domain expertise to improve ontology alignment in the biodiversity field.

Another promising technique for ontology matching is the use of neural networks, which have significantly transformed the landscape by leveraging their ability to understand complex patterns and semantic relationships. In a study by Alexandre Bento et al. [25], convolutional neural networks (CNNs) were used to perform string matching between class labels through character embeddings. The results demonstrated state-of-the-art performance on biomedical ontologies and good performance on non-biomedical ontologies, albeit with some loss of precision. One advantage of this approach is its domain-agnostic nature, making it applicable across various fields. However, challenges remain, such as the need for large amounts of labeled data, potential biases, and the difficulty of maintaining interpretability in complex models. Despite these challenges, the scalability and automatic feature-learning capabilities of neural networks position them as a promising tool for advancing ontology matching.

2.3 Knowledge Graphs

A knowledge graph is a graph-based database designed to represent structured knowledge, enabling precise and efficient data retrieval across interconnected entities and their relationships. It serves as an extensive repository for capturing and organizing complex real-world relationships, facilitating advanced data integration and analysis. Knowledge graphs built upon ontologies leverage these ontologies to define both the structure and semantics of the data, enhancing their ability to connect disparate information sources. By grounding the knowledge graph in an ontology, the underlying structure supports logical reasoning, enabling the inference of new knowledge from existing data.

The construction of a knowledge graph begins with the acquisition and integration of data into a predefined ontological framework. Ontologies, which define the classes, properties, and relationships between concepts, provide the semantic foundation for the graph. Once the data is mapped into an ontology, it is converted into machine-readable formats such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL), which are standards for representing and exchanging knowledge on the semantic web.

In a knowledge graph, each entity is represented as a node, with attributes and properties that describe it. For example, a node representing a "Butterfly" might include attributes such as species name, wing pattern, and habitat preference. Edges between nodes represent relationships between entities, capturing associations such as "is a part of," "is located in," or "has a". These relationships enable rich semantic queries and allow systems to reason over the data. The first step in constructing a knowledge graph is identifying key concepts and entities relevant to the domain, such as individuals, organizations, events, or biological species. Relationships between these entities are then established, creating a comprehensive network of interconnected data points [26].

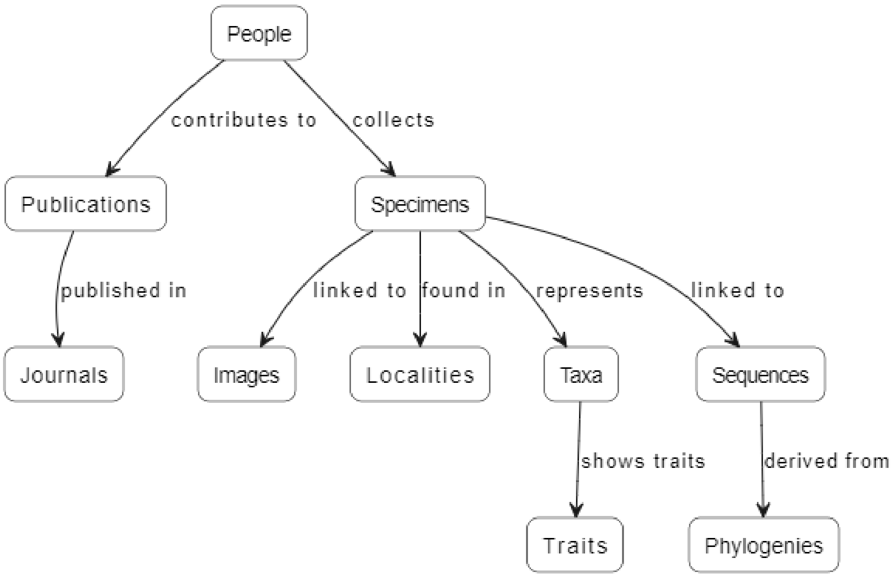


Figure 2.3: A biodiversity data framework connecting specimens, publications, and taxonomic information

After the identification of entities and relationships, they are structured and represented using the "subject-predicate-object" format of the Resource Description Framework (RDF). These triples are then stored in a graph database, which facilitates efficient querying and traversal of relationships [27]. The graph structure, optimized for these operations, enhances the performance of applications that rely on complex data interactions, such as recommendation systems. Knowledge graphs are dynamic, requiring regular updates and maintenance to reflect the addition of new entities and relationships. Over time, as the graph expands, it reveals increasingly sophisticated insights.

By connecting related entities, the Knowledge Graph facilitates exploratory searches, offering deeper and more contextually relevant information. The integration of multiple data sources is essential for building robust knowledge graphs [28]. Scalable reasoning over large datasets is a critical requirement for dynamically generating new knowledge from vast and heterogeneous data sources. In another study [29], researchers discuss how knowledge graphs are being used to enhance machine learning models by enriching them with contextual data, thereby improving their predictive performance. Additionally, advancements in natural language processing (NLP) have been linked to the use of knowledge graphs, particularly in improving tasks such as entity recognition and relationship extraction.

Knowledge graphs significantly enhance information retrieval, entity disambiguation, and provide richer contextual understanding [30]. They provide applications with a structured understanding of entities and their interrelations within a knowledge domain. However, constructing high-quality knowledge graphs is challenging, particularly when integrating data from multiple heterogeneous sources. Ontologies provide the formal foundation by defining concepts, properties, and relationships, ensuring semantic clarity. The integration of ontologies into knowledge graphs enhances the graph's expressiveness, creating a robust system for knowledge representation that supports advanced reasoning and analysis across diverse domains.

Chapter 3

Related Work

This section explores ontology modeling, focusing on the creation and structuring of ontologies. It also covers "Ontology Alignment," which integrates multiple ontologies by identifying correspondences between them. Followed by a discussion on the Biological Collections Ontology (BCO) and Taxrank Ontology.

3.1 Ontology Modeling

The ontology modeling process begins with establishing the scope and purpose of the ontology. This step is crucial as it defines the boundaries and the specific goals that the ontology is intended to achieve within the domain. After defining the scope, the next step is to identify the key concepts that are fundamental to the domain. These concepts are carefully defined to be clear and unambiguous, preventing any misunderstandings or misinterpretations down the line. When dealing with more complex concepts, breaking them down into simpler components can help in creating a more accurate representation. After defining the key concepts, the focus shifts to establishing the relationships among them. These relationships are crucial for illustrating how different entities within the domain are interconnected, contributing to a well-structured and coherent ontology. The next important step in ontology modeling is choosing an appropriate language to represent the concepts and relationships. Two of the most widely used ontology representation languages are the Resource Description Framework (RDF)[1] and the Web Ontology Language (OWL)[2] [31]. RDF organizes data into triples, consisting of a subject (resource), a predicate (property), and an object (value or another resource), creating a hierarchical structure [32]. The flexibility of RDF lies in its ability to describe data without assuming any specific application domain or predefined semantics, making it a versatile tool for various applications.

Building on RDF and RDF Schema (RDFS), OWL offers a more formal and expressive way to define ontologies. OWL is particularly valuable when we need to specify detailed restrictions and cardinality constraints, which allow for precise modeling of complex concepts and relationships. This is especially useful in fields like biodiversity, where the relationships among species, habitats, and ecosystems can be quite intricate. OWL is available in three versions: OWL Lite, OWL DL (Description Logic), and OWL Full [33, 34]. OWL Lite is a simpler version, ideal for basic classification

[1] https://www.w3.org/RDF/

[2] https://www.w3.org/OWL/

tasks. OWL DL is the most popular choice because it strikes a balance between rich modeling capabilities and computational efficiency. OWL Full is the most expressive version, allowing for the greatest flexibility, though it requires more computational resources. By carefully choosing the right ontology language and rigorously defining the concepts and relationships, the ontology modeling process creates a powerful and precise framework. This solid foundation is essential for subsequent tasks like ontology matching and integration.

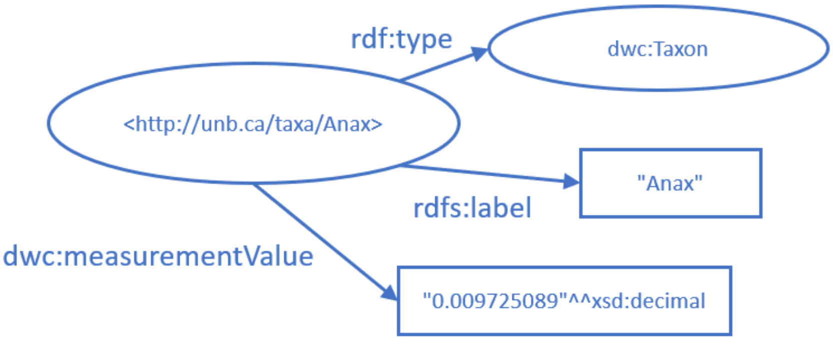


Figure 3.1: RDF graph of biodiversity data [35]

Once the appropriate language for the ontology has been selected, the next step is to implement the ontology using relevant tools and software. This involves converting the conceptual model into machine-readable format that can be utilized by various applications. The complexity of the ontology can vary significantly, depending on the specific requirements of the applications it supports. Essentially, ontologies serve to structure data in a way that accurately reflects the real-world semantics of the entities being represented.

A significant example in biodiversity research is the Biological Collections Ontology (BCO), developed by Walls et al. [36]. Based on the Basic Formal Ontology (BFO) [37], BCO was one of the earliest ontologies created for biodiversity data. It facilitates the semantic exchange of data from various sources, such as museum collections, environmental samples, and ecological surveys. The BCO provides a logical framework that connects samples to data derived from them, enabling better tracking as these samples move through different processes and institutions [38]. This capability is particularly valuable for managing large and complex datasets in biodiversity research.

Over time, a wide array of new ontologies has been developed in the field of biodiversity, each tailored to meet specific needs within the domain. These ontologies are designed to represent different aspects of biodiversity, from species classification and genetic data to ecological interactions and environmental conditions. Despite their shared focus on biodiversity, each ontology is unique in the concepts it represents and the relationships it defines between those concepts. For example, the Lepidoptera Morphology Ontology[3] focuses specifically on the terminology related to the physical characteristics and phenotypes of moths and butterflies. While this ontology is highly relevant to research on Lepidoptera species, its application to the Barcant Butterfly Collection (BBC) is somewhat limited. This is because the BBC dataset pri-

[3] https://obofoundry.org/ontology/lepao.html

marily contains information about the butterflies' location and habitat rather than detailed anatomical data which the ontology represents.

Foundational ontologies like BFO and DOLCE can act as semantic bridges to facilitate the matching of domain-specific ontologies. However, most alignments between foundational and domain ontologies are typically created manually [39, 40].

3.2 Ontology Alignment

The outcome of the ontology matching process is referred to as an ontology alignment. An alignment defines the relationships between entities from distinct ontologies, capturing them with varying levels of precision. These alignments can be used for a wide range of purposes, including ontology merging, query answering, data translation, and navigating the semantic web. It provides a structured set of correspondences between two or more ontologies. This alignment enables the systems involved to communicate effectively, ensuring consistency and coherence across heterogeneous data sources [20]. Ontology alignment is essential for achieving semantic interoperability, allowing independently developed systems and datasets to work together seamlessly. Ontologies, which define structured representations of knowledge through classes, properties, and relationships, often reflect different perspectives and methodologies, especially in complex domains such as healthcare, bioinformatics, and biodiversity. As multiple ontologies emerge across different fields and regions, integration becomes challenging due to differences in structure, scope, and terminology [41].

The goal of ontology alignment is to establish correspondences - such as equivalence, subsumption, or relatedness - between the elements of different ontologies [42]. This enables systems to communicate effectively and analyze data across platforms despite these structural and terminological differences. For example, in healthcare, ontology alignment enables the integration of medical data by mapping clinical terms across different ontologies like SNOMED CT and ICD [43], which are structured differently but represent overlapping information. Similarly, in biodiversity research, aligning ontologies such as the Biological Collections Ontology (BCO) with vocabularies like Darwin Core (DwC) allows species data to be integrated and analyzed across systems and regions.

In the context of the semantic web, [44] ontology alignment connects vast amounts of web data from different services, allowing for more meaningful data exchange between systems that use distinct ontologies. Ontology alignment identifies and maps relationships between entities and concepts in different ontologies, addressing semantic heterogeneity. Automated ontology mapping tools typically leverage both lexical features, such as names and synonyms, and structural features, like relationships between concepts.

When modeling a domain, various levels of complexity need to be considered. These levels range from understanding the meaning and intent behind the words people use when discussing a subject to creating a formal specification of how data is recorded, structured, and exchanged. An ontology modeling technique needs to be able to explain how the terms differ in complexity while highlighting how they can be aligned to infer information across different systems. At the core of this is the concept of "Formal Semantics", which plays a vital role in establishing a clear, precise, and unambiguous interpretation of the intended meaning of concepts. Formal semantics rely on standard vocabularies to define firm relationships between entities, allowing for consistency in interpretation. While links between ontologies are based on actual properties, they can often uncover new or additional information through inference. For instance, if two ontologies share some information about the same concept but one of them includes more detailed attributes or relationships, the formal link between them enables the integration of that additional information. This makes formal semantics particularly useful in cases where two ontologies overlap, as they provide a mechanism to enhance one ontology with information from the other [45].

In practice, this means that by establishing links with formal semantics, ontologies can be expanded or enriched with additional properties or attributes that were not originally present, facilitating deeper insights and more accurate representations of the domain.

3.2.1 Matching Methods

This section provides a overview of the methods used for ontology matching, elaborating on how and when these techniques should be employed. While not an exhaustive list, it outlines key methods adapted from best practices in ontology matching.

Element-Level Matching

Element-level matching techniques focus on individual ontology entities or their instances, without considering their relationships to other entities. These techniques operate at the level of the entity itself, aiming to match similar or related entities across different ontologies [46].

String-Based Matching

This technique relies on comparing the names or descriptions of entities to identify matches. The underlying principle is that entities with similar string patterns are likely to denote the same or related concepts. String-based matching is often a first step in ontology alignment and is widely used because of its simplicity and speed [47].

Linguistic-Based Matching

This approach leverages natural language processing (NLP) tools, lexicons, or domain-specific thesauri to identify relationships between words. It focuses on linguistic properties such as synonymy (different words with the same meaning), homonymy (the same word with different meanings), and partonomy (part-whole relationships). By exploring these linguistic relations, this method improves the accuracy of matching entities, especially in complex domains where the same concept may be described differently [48].

Ontology modeling and alignment involve multiple layers of complexity, from defining the meaning of concepts to specifying how data should be recorded and exchanged. Formal semantics provide the structure for integrating additional information between ontologies, while element-level matching techniques offer practical methods for identifying and aligning entities across different systems. These techniques, when applied correctly, enhance interoperability and allow for more sophisticated knowledge integration across domains [21].

3.2.2 Challenges in Ontology Alignment

- **Semantic Heterogeneity:** Concepts, terms, and relationships are defined differently across ontologies [49]. For instance, the concept of "habitat" may be defined differently, or organisms might be categorized in ways that do not align directly across ontologies. Resolving these differences requires a deep understanding of the context in which the terms are used. Addressing semantic heterogeneity is complex and often resource-intensive, as it requires careful interpretation of domain-specific semantics.

- **Structural Variations:** Ontologies can vary greatly in structure, with some having a flat hierarchy, while others include complex, multi-layered subclass relationships. Aligning these structures requires sophisticated algorithms that can identify similarities across different abstraction levels. The challenge lies in accurately mapping concepts despite these structural differences [50]. For instance, aligning the Darwin Core (DwC) vocabulary, which is relatively flat, with the more detailed Biodiversity Collections Ontology (BCO) requires recognizing not only direct matches but also complex relationships, such as those involving specimen collection methods.

- **Granularity Mismatches:** Granularity mismatches occur when one ontology provides a more detailed or granular representation of a concept compared to another [51]. For instance, an ontology might categorize birds into specific species, whereas another ontology might only have a general "bird" class. This mismatch makes it difficult to map concepts precisely, requiring alignment algorithms to strike a balance between oversimplification and retaining necessary complexity for accurate representation.

- **Context Sensitivity:** The meaning of terms within an ontology can shift based on context. For instance, "habitat" in marine biodiversity may refer to ocean depth or salinity, while in terrestrial biodiversity, it might mean forest type or soil conditions. Accurately aligning such terms requires understanding these contextual differences. Aligning the Marine Metadata Interoperability Ontology with the Environment Ontology (ENVO) would need careful mapping to ensure consistency across marine and terrestrial data.

- **Scalability:** As the number and size of ontologies grow, scaling the alignment process becomes increasingly challenging. Large ontologies with thousands of concepts require significant computational resources, and the complexity of alignment increases exponentially. Developing scalable solutions that can efficiently manage large-scale ontology alignment remains an ongoing research problem [52].

- **Inconsistencies and Conflicts:** After alignment, inconsistencies and conflicts may arise, particularly when integrating data from ontologies with conflicting definitions or relationships. Addressing these inconsistencies often requires establishing formal rules for handling mismatches and prioritizing certain relationships over others [53]. These conflicts are especially pronounced when ontologies represent divergent scientific viewpoints or disciplinary priorities.

3.3 TaxRank and Biological Collections Ontology

In the context of biodiversity informatics and ontology development, TaxRank and the Biological Collections Ontology (BCO) are two crucial ontologies that contribute to organizing and enhancing biodiversity data.

3.3.1 Biological Collections Ontology (BCO)

The Biological Collections Ontology (BCO)[4] plays a crucial role in advancing biodiversity informatics by providing a standardized framework for organizing, integrating, and sharing data related to biological collections. Over the years, many frameworks and ontologies have been developed to improve data management in biodiversity research. Among these, BCO stands out for its ability to connect diverse biological data, making it an essential tool for the organization of biological collections and their associated metadata [54].

BCO is built upon the Basic Formal Ontology (BFO), which offers general concepts like "objects," "qualities," and "processes" that are applicable across various scientific domains. BFO has been widely adopted in fields such as biology, where it serves as a foundation for domain-specific ontologies. Researchers [55] have extensively promoted the use of BFO in biological data representation, and BCO extends this foundational framework to meet the specific needs of biological collections. By incorporating these principles, BCO has become a cornerstone in biodiversity data management, allowing for the integration of complex biological information across various platforms.

The development of ontologies such as ENVO (Environmental Ontology) and OBO Foundry Ontologies laid the groundwork for the representation of biodiversity data. These early efforts provided ways to represent complex environmental data, biological specimens, and collection events [56]. However, they lacked the comprehensive integration capabilities of BCO, which emerged as a more robust framework capable of handling a wider range of data types, including genetic, taxonomic, and ecological information. Early studies in biodiversity informatics were limited to taxonomic classifications or basic environmental parameters, but BCO filled the gap by offering a structured approach to organizing and prioritizing biological collections in a holistic manner.

BCO has been applied in various biological domains, including museum collections, genetic repositories, and ecological surveys. Notable work by Walls et al.[36] demonstrated how BCO facilitates the semantic exchange of data from multiple sources, enabling researchers to effectively track and manage biological samples. This success laid the foundation for other researchers to adopt BCO in their biodiversity related projects [57]. Platforms such as DataONE have integrated BCO to enhance the acces-

[4] https://obofoundry.org/ontology/bco.html

sibility and usability of biodiversity data, further underscoring its importance in managing biological collections. By utilizing BCO, platforms enable researchers around the world to share and query biological collection data more efficiently, streamlining the research process.

In addition to its broad applicability, BCO provides a flexible framework for capturing complex relationships between different types of biological data, such as taxonomic information, genetic sequences, and phenotypic traits. This versatility is crucial in biodiversity studies, where data from diverse sources must be seamlessly integrated to facilitate in-depth ecological and conservation analyses. For example, the Lepidoptera Morphology Ontology focuses on detailed phenotypic data related to butterflies and moths, but its scope is limited to specific use cases, such as morphological studies. In contrast, BCO is capable of accommodating a broader range of data types, including those related to collection events, environmental conditions, and genetic sequences, making it a more comprehensive solution for biodiversity research [58].

Building on this foundation, BCO is applied to the Barcant Butterfly Collection (BBC) to transform the valuable dataset into a machine-readable format. By using BCO, the collection's metadata—such as specimen data, collection events, and environmental conditions—are well-structured and semantically linked. This enables better querying, cross-referencing, and integration with other biodiversity datasets. By leveraging BCO's capacity to capture complex data relationships, this research aims to enhance the accessibility and usability of the BBC dataset for researchers engaged in biodiversity and conservation efforts. The decision to use BCO for this project stems from its adaptability and scalability in handling various types of biodiversity data. Unlike more domain-specific ontologies, which may focus solely on taxonomy or environmental data, BCO provides a flexible and comprehensive solution for managing biological collections and their associated processes. This adaptability is crucial for the BBC, a large dataset containing thousands of specimens, each with unique attributes such as habitat, geographical location, and collection details. BCO's foundation on BFO also ensures interoperability across platforms, an essential feature in biodiversity informatics. Additionally, BCO has a proven track record in large biodiversity projects, giving confidence that it will effectively handle the BBC dataset.

The research adapts the BCO framework to specifically address the needs of butterfly collections. It incorporates attributes such as habitat data, geographical locations, and collection event specifics, which provide a more focused framework for butterfly research. This customization allows for more accurate data representation, making the BBC dataset more valuable to researchers studying butterflies. Furthermore, it ensures that the BBC dataset, once structured using BCO, is interoperable with major biodiversity platforms like GBIF and DataONE, facilitating easier access and use by researchers across multiple fields, including ecology, genetics, and taxonomy. By transforming the Barcant Butterfly Collection (BBC) dataset into a machine-readable format using BCO, this effort not only improves its accessibility but also enhances its potential for conservation efforts and research prioritization. This transformation will enable researchers to identify species at risk, study migration patterns, and analyze the effects of environmental changes on butterfly populations. Additionally, this work contributes to the broader field of biodiversity informatics by demonstrating how BCO can be applied to other biological collections, providing a framework that can be replicated or adapted by other institutions seeking to improve their management of biodiversity data.

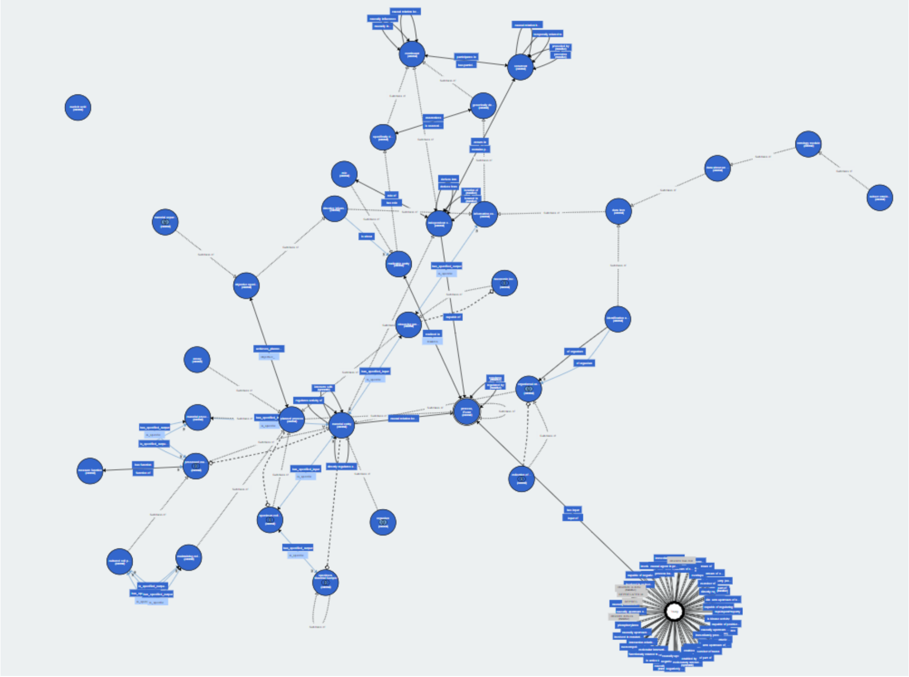


Figure 3.2: Visualization of the BCO Ontology

3.3.2 TaxRank

Taxonomic classification has long been central to biodiversity research, offering a systematic way to organize species based on their evolutionary relationships and characteristics. However, as biodiversity datasets grow in complexity, traditional methods of classification are no longer sufficient to address the growing need to prioritize species for research, conservation, and ecological management. To meet these needs, new methods such as Phylogenetic Diversity (PD) and Evolutionary Distinctiveness (ED) have been developed, which rank species based on their evolutionary significance within the phylogenetic tree [59]. These approaches allow to identify species that play key roles in evolutionary history, guiding efforts to conserve those that are most critical to preserving biodiversity.

In the field of conservation, prioritization has become a key challenge. Tools like the Species Prioritization Index (SPI) are specifically designed to identify species at the greatest risk of extinction or those that play vital ecological roles. However, these tools are limited to conservation-focused contexts. This is where TaxRank, a method for ranking species by their taxonomic relevance within large biodiversity datasets, becomes essential. TaxRank extends beyond conservation, offering a comprehensive way to prioritize species across different research domains. Its ability to filter and rank species by various criteria helps researchers and conservationists target the most relevant species, streamlining decision-making processes in biodiversity management [60].

While ontologies such as the Environment Ontology (ENVO) and the Biological

Collections Ontology (BCO) have advanced the structuring and sharing of biodiversity data, they lack built-in mechanisms for ranking species based on their importance. These ontologies provide a framework for connecting diverse datasets but do not emphasize the prioritization of species according to their taxonomic or ecological relevance [61]. TaxRank addresses this gap by introducing a ranking system within the framework of these ontologies, ensuring that significant species are highlighted in biodiversity research. By incorporating a ranking mechanism into ontology-driven databases, TaxRank enhances the utility of these tools, providing a more targeted approach to managing biodiversity information [62]. In the context of the semantic web and biodiversity knowledge graphs, which are designed to integrate taxonomic data across multiple platforms, the sheer volume of species data can be overwhelming [63]. While knowledge graphs facilitate data interoperability, they can make it difficult for researchers to focus on the most critical species. TaxRank plays a crucial role in managing complexity by introducing a method for prioritizing taxa within these interconnected systems. TaxRank ensures that the most important species are brought to the forefront of analysis. This integration of TaxRank into semantic technologies and ontologies not only enhances the organization of biodiversity data but also strengthens efforts to conserve and study the species that are most vital to ecological and evolutionary processes.

The decision to focus on TaxRank stems from its scalability and adaptability, which align with the objectives of managing and prioritizing species within extensive biodiversity datasets, such as the Barcant Butterfly Collection (BBC). TaxRank's inherent flexibility in ranking taxa based on variouus criteria such as relevance, ecological importance, or conservation status makes it particularly useful for handling datasets that encompass thousands of species. Its ability to integrate seamlessly with biodiversity ontologies allows for more sophisticated querying and enhanced data interoperability, positioning it for large-scale biodiversity research.

The TaxRank methodology is being refined to better address specific attributes of biodiversity datasets, such as species vulnerability, geographical significance, and ecological roles. This will enhance the system's ability to provide precise prioritization, particularly in datasets like the BBC, which are critical for conservation strategies and research initiatives. The methodology is also integrated with existing biodiversity ontologies, such as the Biological Collections Ontology (BCO), enhancing interoperability and streamlined data management. TaxRank's applicability to biodiversity knowledge graphs is expanded, ensuring that ranked datasets are accessible and usable within semantic web environments. This integration with knowledge graphs enhances TaxRank's role in biodiversity research by making data more usable and relevant for large-scale studies. Automated processes within TaxRank are developed to scale the methodology for handling large datasets, reducing manual workload and ensuring accurate and efficient taxonomic data management.

Chapter 4

Dataset

4.1 Dataset Description

The primary dataset for this research is the Barcant Butterfly Collection (BBC), sourced from the Global Biodiversity Information Facility (GBIF)[1], a global platform for aggregating and analyzing biodiversity data. GBIF provides a vast array of records, from historical museum specimens dating back centuries to modern DNA barcodes and digital images. This platform, supported by contributions from over 107 countries, serves as a comprehensive repository for species occurrence and distribution data worldwide.

The BBC is the largest butterfly collection in the Caribbean, with over 5,000 specimens representing nearly 700 species native to Trinidad and Tobago. This collection was curated over 50 years by Malcolm Barcant, a Trinidadian entomologist, and is thoroughly documented in his work, Butterflies of Trinidad and Tobago [64]. Housed at the Angostura Museum and Butterfly Collection in Port of Spain since its acquisition by Angostura in 1974[2], the BBC offers invaluable insights into the region's butterfly fauna, highlighting many rare and endemic species that play essential roles in ecosystem health and biodiversity.

With support from the University of the West Indies Zoology Museum (UWIZM), the collection was digitized, enabling the transcription of specimen data from handwritten labels for accessibility on the GBIF portal. The BBC is highly valued in scientific circles for its detailed documentation, including field notes, photographs, and records on butterfly morphology, coloration, and patterns. This makes it a crucial resource for taxonomic and ecological research, offering an extensive and well-preserved snapshot of Trinidad and Tobago's butterfly diversity.

4.2 Significance of the Collection in Biodiversity Research

As a comprehensive record of butterfly species in Trinidad and Tobago, the collection provides essential baseline data that help researchers monitor changes in species diversity and distribution over time. Such data are critical for assessing the impact of environmental changes and human activities on local biodiversity [7]. The collection's significance is further underscored by its contributions to taxonomic studies.

[1] https://www.gbif.org

[2] https://www.angostura.com/tours

Specimens from the Barcant Collection have been instrumental in describing new species and refining the classification of existing ones, offering valuable insights into the evolutionary relationships among butterfly species. The collection contributed to the revision of species within the Heliconiini and Morphinae groups of butterflies, leading to the identification of new species and a better understanding of their phylogeny. Additionally, studies using the Barcant Collection have provided essential data for constructing evolutionary trees, offering insights into how butterfly species have diversified over time. Research publications from entomologists like Norman C. Owen [10] and related papers on Neotropical butterfly systematics often reference the Barcant Collection in their work. The historical nature of the collection allows researchers to compare past and present species distributions. Recently, efforts have been made to digitize the Barcant Collection, making it accessible to researchers worldwide. This digital availability has expanded the collection's impact, facilitating comparative studies across different regions and promoting collaborative research in biodiversity science.

The application of ontologies and knowledge graphs can significantly enhance data management and analysis of the Barcant Butterfly Collection. Ontologies provide a standardized framework for describing butterfly morphology, taxonomy, and distribution, enhancing data accuracy and clarity. Knowledge graphs, built on these ontologies, create interconnected networks that reveal complex relationships among species, enabling advanced analyses and helps uncover hidden patterns [65]. This integration boosts the accessibility and utility of the digitized collection, supporting more effective and collaborative research.

4.3 Structure of the Dataset

The collection is thoughtfully categorized into six distinct family groups:

1. **Hesperiidae**[3] - Known as skippers, these small, fast-flying butterflies are distinguished by their hooked antennae. Notable species include the Long-tailed Skipper and the Red-banded Hairstreak.

2. **Pieridae**[4] - This family is renowned for its brightly colored members such as the Cloudless Sulphur and Orange-barred Sulphur, which are common in the region.

3. **Nymphalidae**[5] - The largest butterfly family, often referred to as brush-footed butterflies, includes species like the Blue Emperor and the Malachite. Nymphalidae are known for their striking wing patterns and colors.

4. **Papilionidae**[6] - Some of the largest butterflies, commonly known as Swallowtails, characterized by their tail-like extensions on the hindwings. The Zebra Longwing is a prominent species in this family.

[3] https://www.gbif.org/dataset/a4b2035f-fa9d-4d80-97dc-f5d58ec4ef51

[4] https://www.gbif.org/dataset/438245ee-99e3-4417-bd4f-4ef7ecc16660

[5] https://www.gbif.org/dataset/641196ce-d154-456a-8af1-a306b0f81895

[6] https://www.gbif.org/dataset/a65807ea-c9e7-4173-8ed2-8c41e1b3a0fe

5. **Lycaenidae**[7] - These small, often iridescent butterflies include species such as Clench's Greenstreak and Cassius Blue, highlighting the delicate beauty of the collection.

6. **Riodinidae**[8] - Often called Metalmarks due to their metallic wing markings, this family includes species like the Metallic-tipped Flasher and the Doris Longwing, adding to the visual diversity of the collection.

Family	**Notable Species**
Hesperiidae	Long-tailed Skipper, Red-banded Hairstreak
Pieridae	Cloudless Sulphur, Orange-barred Sulphur
Nymphalidae	Blue Emperor, Malachite
Papilionidae	Zebra Longwing
Lycaenidae	Clench's Greenstreak, Cassius Blue
Riodinidae	Metallic-tipped Flasher, Doris Longwing

Table 4.1: Notable Butterfly Species by Family

Each entry in the dataset contains comprehensive information about specific butterfly specimens. The attributes for each entry are as follows:

- **occurrenceID**: A unique identifier for each specimen occurrence.
- **scientificName**: The full scientific name of the butterfly
- **taxonRank**: The taxonomic rank of the most specific name in the scientificName.
- **decimalLatitude** and **decimalLongitude**: The geographic coordinates where the specimen was recorded.
- **identifiedBy**: The name of the person who identified the specimen.
- **dateIdentified**: The date when the identification was made.
- **eventDate**: The date when the specimen was collected, defining the temporal aspect of occurrences.
- **locality**: The specific locality where the specimen was collected, providing detailed location context.
- **family**: The taxonomic family to which the organism belongs, aiding in hierarchical classification.
- **genus**: The genus within the taxonomic hierarchy, further refining the classification.
- **species**: The specific epithet within the genus, representing the organism at the species level.

[7] https://www.gbif.org/dataset/2e8e59ba-4132-47ab-9923-dfc0c046683b

[8] https://www.gbif.org/dataset/8497849e-773d-4d4b-b89c-221166879406

4.4 Data Visualization

Data visualization plays a crucial role in simplifying complex data, revealing patterns, and improving retention. It supports predictive analysis, enhances storytelling, and increases productivity by encouraging exploration. Transforming large datasets into visual formats makes information more understandable and engaging, enabling quick insight generation and informed decision-making.

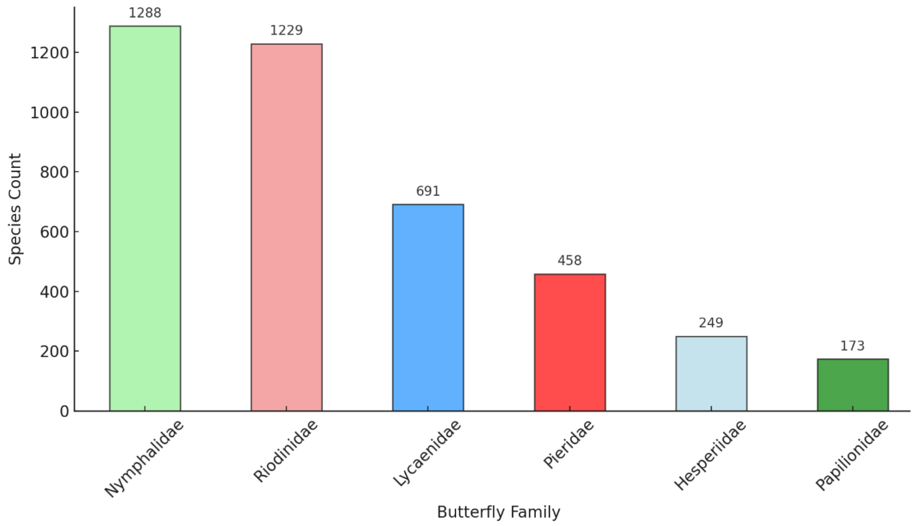


Figure 4.1: Species count per family

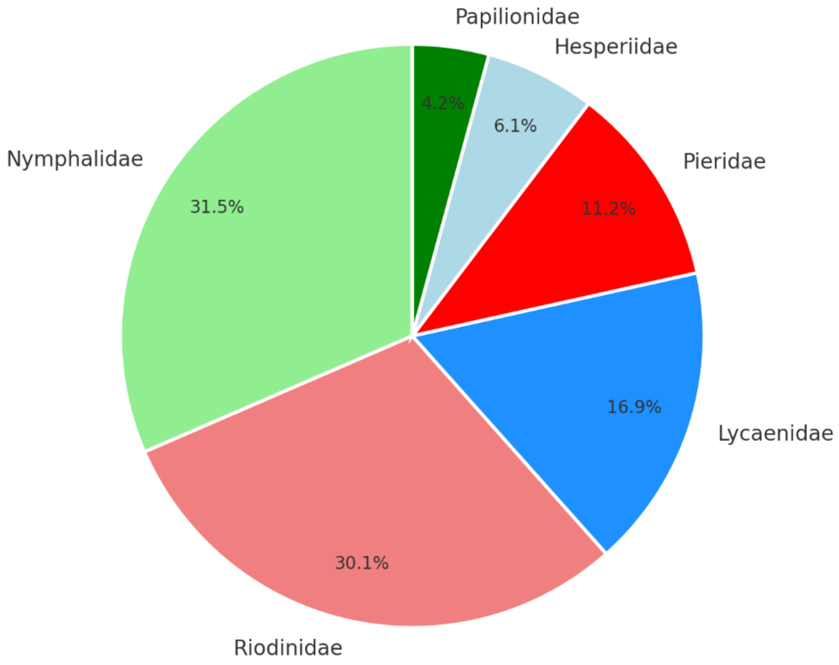


Figure 4.2: Proportion of Species by Butterfly Family

The bar and pie charts in fig. 4.1 and fig. 4.2 provide a detailed visualization of species distribution among butterfly families, highlighting both numerical counts and proportional significance. Nymphalidae stands out as the most dominant family, representing 31.5% of the total species with 1,288 recorded individuals, while Papilionidae shows the smallest count among the six families, with 173 species. This dominance of Nymphalidae aligns with its broad ecological range and adaptability across diverse habitats, from tropical to temperate ecosystems, affirming its recognized importance in biodiversity studies. The combined graphical representation underscores the critical role that Nymphalidae plays in butterfly biodiversity within the dataset, while also highlighting the contributions of other families such as Riodinidae, Lycaenidae, and Pieridae.

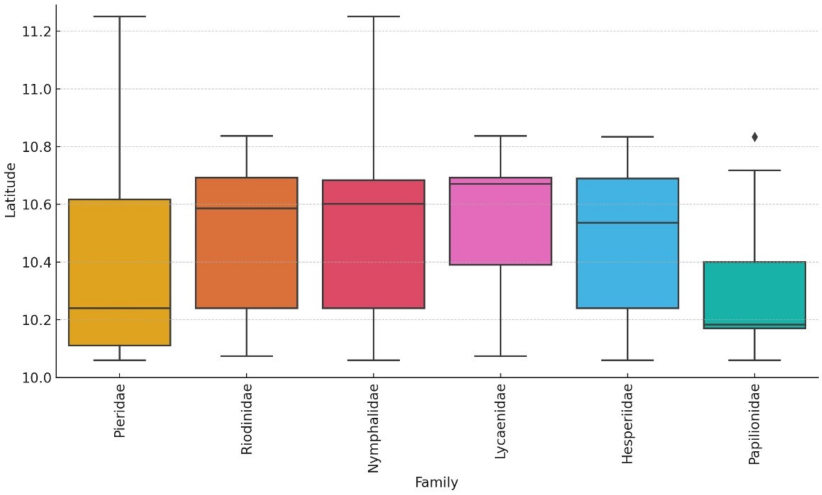


Figure 4.3: Latitudinal distribution of species occurrences across butterfly families

The boxplot in fig. 4.3 represents the distribution of species occurrence by family based on latitude. Six families are depicted: Pieridae, Riodinidae, Nymphalidae, Lycaenidae, Hesperiidae, and Papilionidae. Each box plot displays the interquartile range (IQR), median, and spread of latitudes for species occurrences within each family. The species in the Pieridae family demonstrate the largest range in latitude, with some data points extending beyond 11.2. Conversely, Papilionidae species have a more constrained distribution. Outliers are present, particularly for Hesperiidae, suggesting some variability in species occurrences. This visualization helps highlight the latitudinal distribution and variability of different butterfly families.

Upon analysis of the line graph fig. 4.4, it is evident that the dataset is well-structured, exhibiting minimal instances of missing or null values across the considered columns. The line graph depicts the number of species names against the proportion of null values for each of the six distinct butterfly families. The family Pieridae notably has the highest occurrence of null values, indicating a potential data gap that warrants closer scrutiny.

In contrast, the Riodinidae family presents a notable anomaly within the dataset, characterized by the inclusion of 349 null values in the family column and the presence of other families within the same dataset. This atypical observation is captured and highlighted in the accompanying bar chart fig. 4.5.

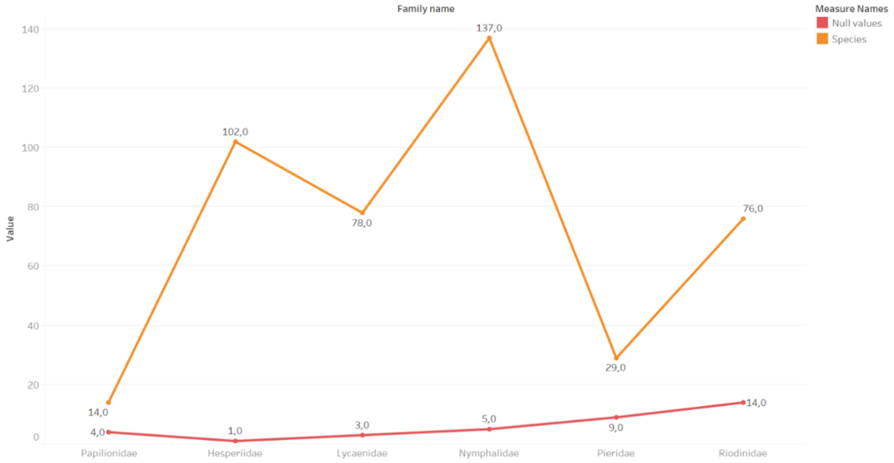


Figure 4.4: Trend of null values with respect to species count

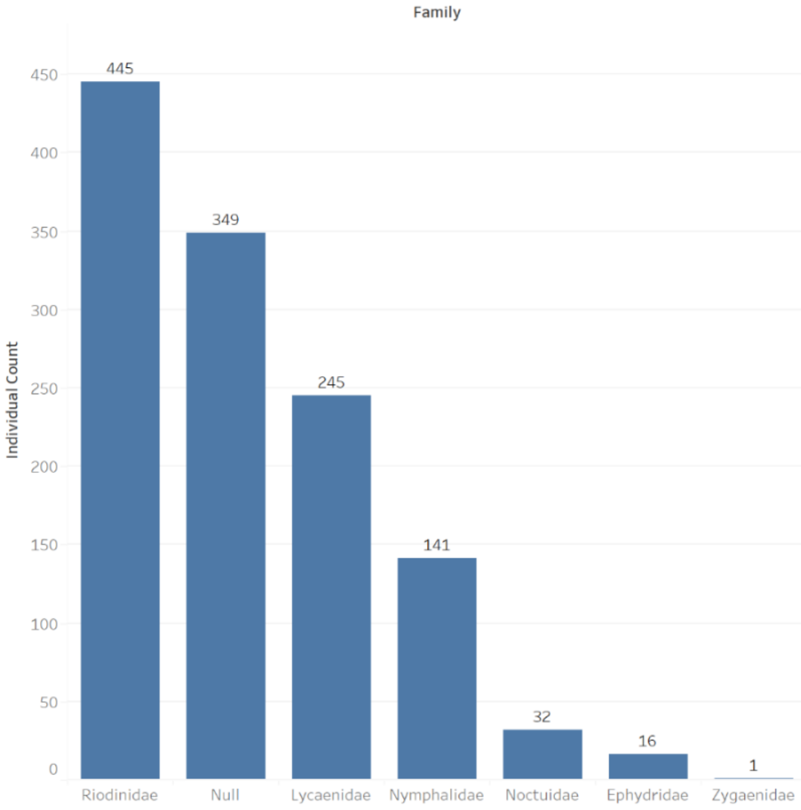


Figure 4.5: Riodinidae collection with count

Chapter 5

Methodology

This chapter provides a comprehensive overview of the research methodologies and strategies used to address the main research question at hand:

> **RQ1**: *Concepts and Relationships: What specific ontological concepts and relationships are crucial for developing a comprehensive Barcant Butterfly Collection ontology that effectively captures the nuances of butterfly biodiversity?*

The following serves as the chapter's framework:

1. Transform the Barcant Butterfly Collection (BBC) dataset into the Darwin Core (DwC) vocabulary to standardize biodiversity data. This ensures that the data follows a widely-accepted format for biodiversity information.

2. Convert the standardized DwC data into RDF (Resource Description Framework) format to enhance data interoperability and enable advanced querying capabilities. This process can be done by creating and designing ontology for the dataset

3. Apply ontology alignment techniques to match the created BBC ontology with existing biodiversity ontologies, such as the Biological Collections Ontology (BCO) and the Taxonomic Rank Ontology (TaxRank).

4. Construct a knowledge graph using the aligned ontology data for advanced analysis, enabling more in-depth exploration and querying of the BBC dataset within a broader ecological and environmental context.

5. Ecological range maps are generated to visualize the geographic distribution of butterfly species in the BBC dataset, offering insights into species habitats and spatial patterns. These maps are crucial for conservation planning, as they help identify areas of species richness and regions where biodiversity may be at risk.

6. Compare the Barcant Butterfly Collection (BBC) with the World Database on Protected Areas (WDPA), with focus on identifying butterfly species from the BBC that are found within the geographic boundaries of protected areas listed in the WDPA. This will enrich the present dataset and can aid conservation strategies

5.1 Data Transformation and Standardization to Darwin Core (DwC)

To ensure that the dataset is compatible with other biodiversity datasets and can be easily integrated into global databases, it is essential to map the attributes to the Darwin Core (DwC) vocabulary. Darwin Core is a standardized framework used widely for representing biodiversity data. By mapping the dataset's attributes to DwC terms, the data becomes more accessible and interoperable with other datasets in the scientific community. It allows the data to be shared, compared, and analyzed alongside other datasets from different regions or projects. This standardization is crucial for large-scale biodiversity assessments, species distribution modeling, and conservation planning. Based on the attributes, the Darwin Core terms for BBC dataset are given as:

- scientificName → dwc:scientificName
- locality → dwc:locality
- eventDate → dwc:eventDate
- recordedBy → dwc:recordedBy
- decimalLatitude → dwc:decimalLatitude
- decimalLongitude → dwc:decimalLongitude
- coordinateUncertaintyInMeters → dwc:coordinateUncertaintyInMeters
- countryCode → dwc:country
- verbatimScientificName → dwc:verbatimScientificName
- basisOfRecord → dwc:basisOfRecord
- institutionCode → dwc:institutionCode
- countryCode → dwc:country
- collectionCode → dwc:collectionCode
- catalogNumber → dwc:catalogNumber
- identifiedBy → dwc:identifiedBy
- dateIdentified → dwc:dateIdentified

5.2 Ontology Construction

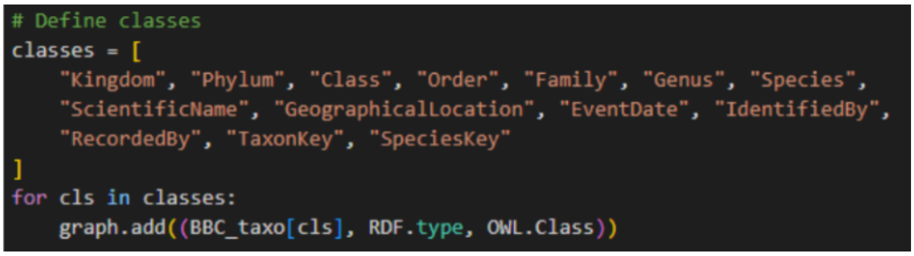
The ontology for the Barcant Butterfly Collection (BBC) is meticulously constructed using RDF and OWL standards, focusing on the integration and standardization of diverse datasets, each with unique aspects. This section outlines the ontology creation process, highlighting its design, application, classes, properties, and the mapping of entities to ensure compatibility and interoperability with existing ontologies in the domain of biodiversity research.

5.2.1 Ontology Design

The formal ontology framework presented in this research involves creating structured classes, properties, and relationships. The ontology enables the systematic representation of butterfly species, their taxonomic classification, geographic locations, and observation data.

1. **Classes**
   The ontology is composed of several core classes, which represent taxonomic ranks and geographic locations. These classes ensure that the ontology can capture both biological and geospatial information related to butterfly observations. Each class is designed to capture specific information, corresponding to the unique characteristics of the datasets.

   - **Taxonomy**: The Taxonomy class organizes the taxonomic hierarchy with subclasses such as Kingdom, Phylum, Class, Order, Family, Genus, and Species. Each of these classes follows a hierarchical structure that ensures the accurate classification of butterfly species from the most general (Kingdom) to the most specific (Species).
   - **ScientificName**: A class representing the scientific name of a species. This class is essential for species identification following the binomial nomenclature system
   - **GeographicLocation**: This class captures the spatial data of observations, including latitude and longitude. The geographic location class is used to record where a butterfly species has been observed.
   - **Observation**: Represents data about specimen observations, such as the date of observation and the observer.



```
# Define classes
classes = [
    "Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species",
    "ScientificName", "GeographicalLocation", "EventDate", "IdentifiedBy",
    "RecordedBy", "TaxonKey", "SpeciesKey"
]
for cls in classes:
    graph.add((BBC_taxo[cls], RDF.type, OWL.Class))
```

Figure 5.1: Diagram of ontology structure

2. **Properties**
   The properties in the ontology define the relationships between the taxonomic categories, geographic locations, and observational data. These properties are essential for linking instances of classes and ensuring that the ontology captures the full complexity of butterfly species data.

   - **hasScientificName**: This object property links a Species instance to its scientific name, represented in the ScientificName class. This property ensures the formal identification of species within the ontology.

- **observedIn**: This property links an instance of the Species class to an instance of the GeographicLocation class, indicating the specific location where the butterfly was observed.
- **Taxonomic Relationships**: Properties such as hasPhylum, hasClass, hasOrder, hasFamily, hasGenus, and hasSpecies are defined to link species to their respective taxonomic categories, ensuring the accurate representation of the biological hierarchy.

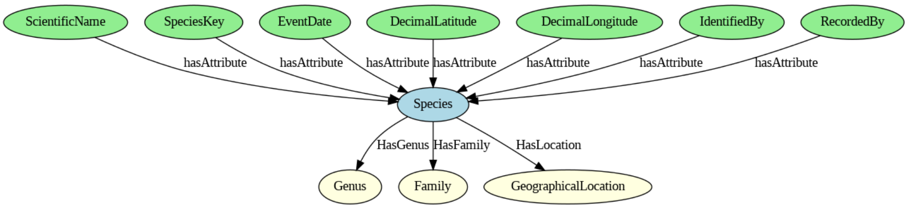
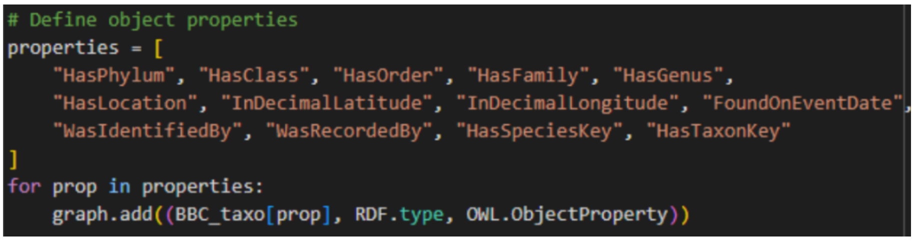


Figure 5.2: View of the Species Entity in the BBC Ontology

Figure 5.2 presents a detailed diagram of the "Species" entity in the Barcant Butterfly Collection ontology. In this figure, the Species entity realizes the core function related to a set of attributes and related entities via specific relationships.

- HasGenus: Links the species to its Genus, situating it within the broader taxonomic hierarchy.
- HasFamily: Connects the species to its Family, further defining its taxonomic classification.
- HasLocation: Relates the species to its GeographicalLocation, integrating geographical context into the species profile.



```
# Define object properties
properties = [
    "HasPhylum", "HasClass", "HasOrder", "HasFamily", "HasGenus",
    "HasLocation", "InDecimalLatitude", "InDecimalLongitude", "FoundOnEventDate",
    "WasIdentifiedBy", "WasRecordedBy", "HasSpeciesKey", "HasTaxonKey"
]
for prop in properties:
    graph.add((BBC_taxo[prop], RDF.type, OWL.ObjectProperty))
```

3. **Hierarchy**

The taxonomic hierarchy is a critical part of ontology, ensuring that species are correctly classified according to the biological system. The `isSubclassOf` property is used to establish this hierarchy, linking broader taxonomic categories to more specific ones. This property defines the relationship between taxonomic classes. For example, the `Species` class is a subclass of the `Genus` class, the `Genus` class is a subclass of the `Family` class, and so on. This hierarchical structure preserves the natural biological classification of butterfly species.

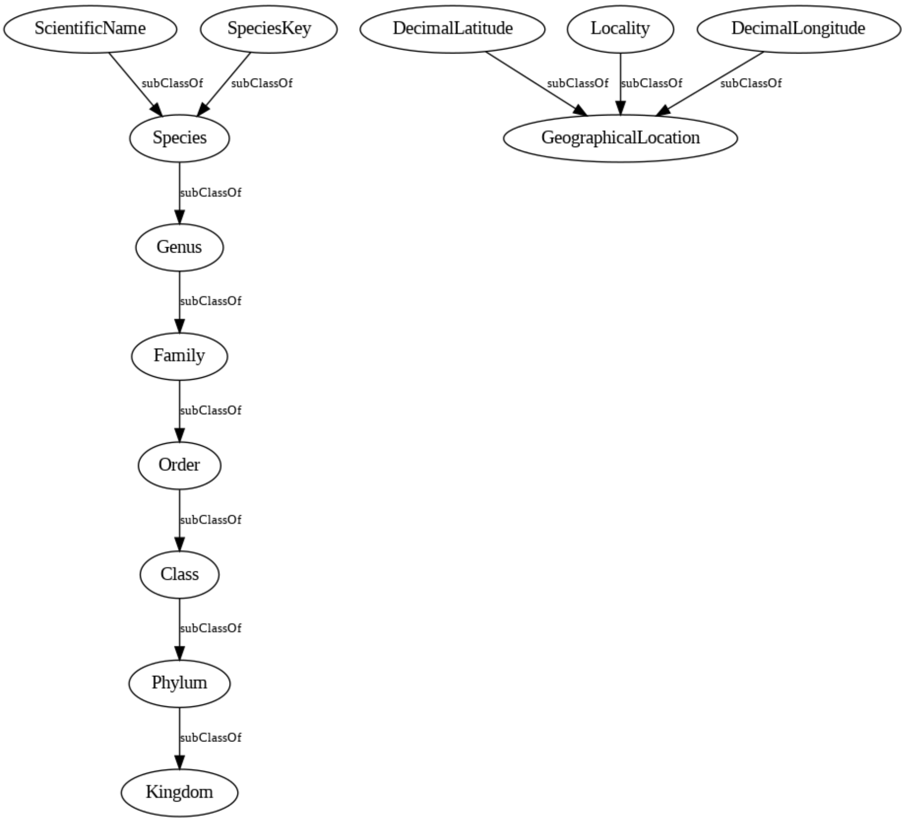
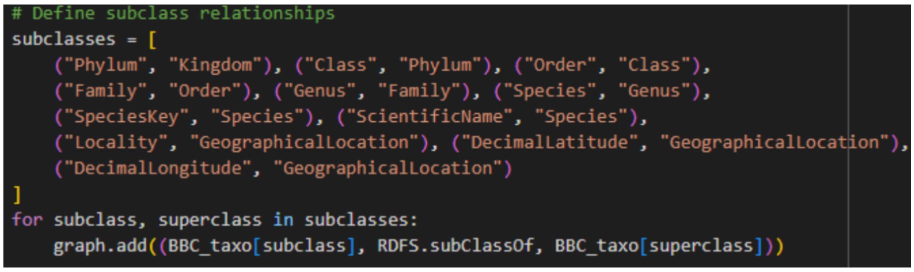


Figure 5.3: Ontology Structure Diagram of the Barcant Butterfly Collection

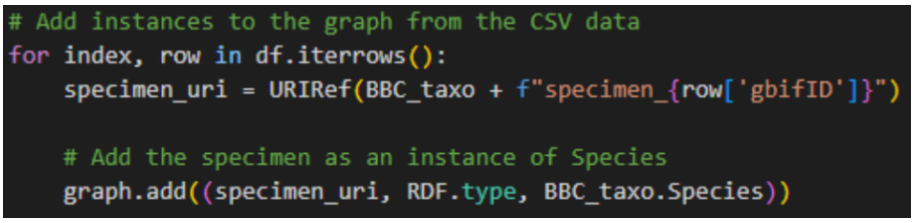
The hierarchical structure of the ontology is depicted in fig. 5.3, illustrating how the main classes and their subclasses are organized. This diagram serves as a foundational visual representation of the ontology, demonstrating how it categorizes and organizes taxonomic information. It facilitates accurate data representation and supports complex queries and data integration.



```
# Define subclass relationships
subclasses = [
    ("Phylum", "Kingdom"), ("Class", "Phylum"), ("Order", "Class"),
    ("Family", "Order"), ("Genus", "Family"), ("Species", "Genus"),
    ("SpeciesKey", "Species"), ("ScientificName", "Species"),
    ("Locality", "GeographicalLocation"), ("DecimalLatitude", "GeographicalLocation"),
    ("DecimalLongitude", "GeographicalLocation")
]
for subclass, superclass in subclasses:
    graph.add((BBC_taxo[subclass], RDFS.subClassOf, BBC_taxo[superclass]))
```

4. **Instances**
   Instances in the ontology correspond to specific data points, including individual butterfly species, geographic locations, and observations. These instances are essential for populating the ontology with real-world data, which enables researchers to analyze species distribution and track biodiversity trends. Each butterfly observation is instantiated as a member of the Species class, with connections to relevant taxonomic classes and geographic information. Each instance is assigned a unique URI, often sourced from external datasets like the Global Biodiversity Information Facility (GBIF), ensuring precise identification

and integration.



```
# Add instances to the graph from the CSV data
for index, row in df.iterrows():
    specimen_uri = URIRef(BBC_taxo + f"specimen_{row['gbifID']}")

    # Add the specimen as an instance of Species
    graph.add((specimen_uri, RDF.type, BBC_taxo.Species))
```

5.2.2 Ontology Population

Populating the ontology is a crucial step in turning a conceptual model into a data-rich framework that can be used for real-world analysis. In this case, the ontology is populated with biodiversity data related to butterfly species, including taxonomic classification, geographic location, and specimen observation information. The data is drawn from a CSV file that contains various attributes for each observed specimen. The ontology used in the research project is a custom biodiversity ontology focused on representing taxonomic classifications and specimen observations of butterfly species. It's built using RDF (Resource Description Framework) and OWL (Web Ontology Language) to organize biodiversity data.

Figure 5.4 represents a biodiversity ontology designed to manage butterfly specimen data, focusing on taxonomic classification, geographic location, and observation details. The `ButterflySpecimen` class captures individual butterfly data, including its unique identifier, scientific name, and metadata such as geographic coordinates and observation dates. Each specimen is linked to a hierarchical Taxonomy class, which categorizes it under biological ranks such as Kingdom, Phylum, and Species. The `GeographicLocation` class holds information about where the specimen was observed, while the `Taxonomy` class provides external mappings to the widely-used NCBI taxonomy database, ensuring data interoperability. The QueryService interface supports data queries, allowing users to retrieve information about specimens, their classifications, locations, and observation metadata, which helps facilitate biodiversity research and analysis. The process of populating the ontology involves creating individual instances for each specimen and linking them to relevant properties, such as their position in the biological taxonomy, the location where they were observed, and other relevant metadata (e.g., the observer and the date of identification).

1. **Instance Creation**
   The .csv file contains detailed information about butterfly specimens, including attributes like gbifID, phylum, class, order, family, genus, and geographic coordinates such as decimalLatitude and decimalLongitude. For each row in the CSV file, the RDF graph creates an instance of the Species class, using the gbifID as a unique identifier for the specimen. The URI for each specimen instance is dynamically generated based on its gbifID to ensure each instance is uniquely represented in the graph.

2. **Property Assignment**
   Once the instance for each specimen is created, various properties are assigned

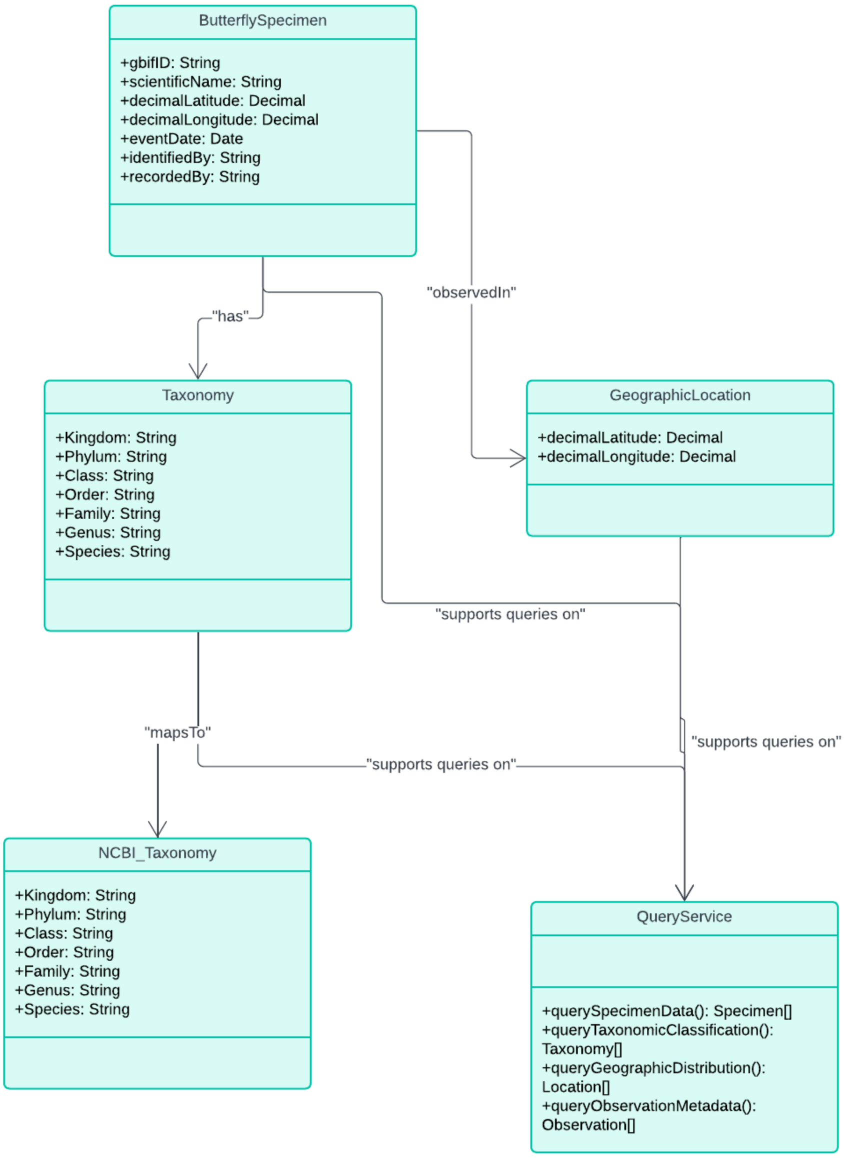
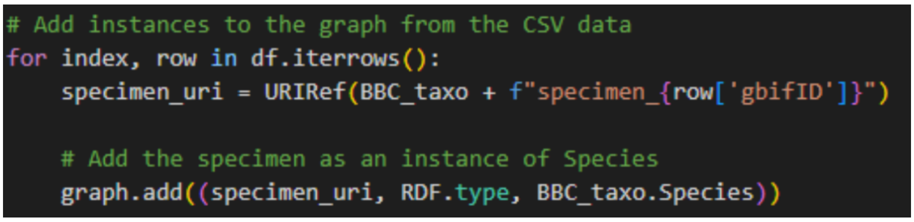


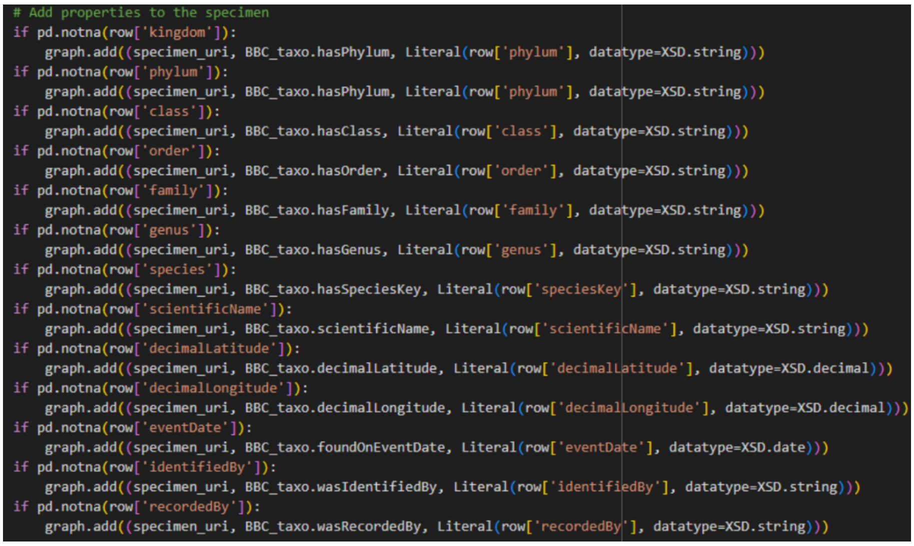
Figure 5.4: Class diagram of the ontology used



```
# Add instances to the graph from the CSV data
for index, row in df.iterrows():
    specimen_uri = URIRef(BBC_taxo + f"specimen_{row['gbifID']}")

    # Add the specimen as an instance of Species
    graph.add((specimen_uri, RDF.type, BBC_taxo.Species))
```

to it based on the data available in the CSV file. These properties include both taxonomic information (e.g., Phylum, Class, Order) and geographical information (e.g., decimalLatitude, decimalLongitude). Each property is only added if there is data present for it in the CSV file, which is checked using the pd.notna() function to avoid adding empty or missing values.



```
# Add properties to the specimen
if pd.notna(row['kingdom']):
    graph.add((specimen_uri, BBC_taxo.hasPhylum, Literal(row['phylum'], datatype=XSD.string)))
if pd.notna(row['phylum']):
    graph.add((specimen_uri, BBC_taxo.hasPhylum, Literal(row['phylum'], datatype=XSD.string)))
if pd.notna(row['class']):
    graph.add((specimen_uri, BBC_taxo.hasClass, Literal(row['class'], datatype=XSD.string)))
if pd.notna(row['order']):
    graph.add((specimen_uri, BBC_taxo.hasOrder, Literal(row['order'], datatype=XSD.string)))
if pd.notna(row['family']):
    graph.add((specimen_uri, BBC_taxo.hasFamily, Literal(row['family'], datatype=XSD.string)))
if pd.notna(row['genus']):
    graph.add((specimen_uri, BBC_taxo.hasGenus, Literal(row['genus'], datatype=XSD.string)))
if pd.notna(row['species']):
    graph.add((specimen_uri, BBC_taxo.hasSpeciesKey, Literal(row['speciesKey'], datatype=XSD.string)))
if pd.notna(row['scientificName']):
    graph.add((specimen_uri, BBC_taxo.scientificName, Literal(row['scientificName'], datatype=XSD.string)))
if pd.notna(row['decimalLatitude']):
    graph.add((specimen_uri, BBC_taxo.decimalLatitude, Literal(row['decimalLatitude'], datatype=XSD.decimal)))
if pd.notna(row['decimalLongitude']):
    graph.add((specimen_uri, BBC_taxo.decimalLongitude, Literal(row['decimalLongitude'], datatype=XSD.decimal)))
if pd.notna(row['eventDate']):
    graph.add((specimen_uri, BBC_taxo.foundOnEventDate, Literal(row['eventDate'], datatype=XSD.date)))
if pd.notna(row['identifiedBy']):
    graph.add((specimen_uri, BBC_taxo.wasIdentifiedBy, Literal(row['identifiedBy'], datatype=XSD.string)))
if pd.notna(row['recordedBy']):
    graph.add((specimen_uri, BBC_taxo.wasRecordedBy, Literal(row['recordedBy'], datatype=XSD.string)))
```

5.3 Ontology Alignment

Ontology alignment is a critical step in integrating multiple ontologies or datasets, ensuring that equivalent or similar concepts are linked across different sources while maintaining semantic consistency. The primary goal of aligning the BBC Ontology (Barcant Butterfly Collection Ontology) with the Biological Collections Ontology (BCO) and the Taxonomic Rank Ontology (TaxRank) is to facilitate seamless integration of taxonomic, biological collection, and observational data across diverse sources. This alignment enhances interoperability, allowing compatibility with existing biodiversity data infrastructures and enabling streamlined data exchange and reuse across institutions. By linking specimen observations from BBC with taxonomic information from TaxRank and specimen-level data from BCO, the dataset is enriched. Consistency is achieved by mapping equivalent classes and properties, ensuring uniform representation of key concepts, which simplifies data retrieval and integration. Additionally, the alignment process strengthens the ontology's querying and reasoning capabilities, enabling sophisticated queries that span taxonomic, specimen collection, and observational data. This comprehensive approach supports interoperability and advanced reasoning in biodiversity informatics.

The Biological Collections Ontology (BCO) is a specialized framework designed to standardize the representation of data associated with biological collections, including specimen information, collection events, taxonomic details, and the preservation and curation of specimens. It is widely used by institutions like natural history museums, herbaria, and biological repositories to ensure consistency in how they manage and share information about their collections. BCO enables the structured documentation of specimen data, encompassing key attributes such as taxonomic classification (species, genus, family), physical traits, and scientific names. Additionally, it captures critical information about collection events, including the date and location of specimen collection, the individuals involved, and the methods used. The ontology also provides a structured means of representing taxonomic information, ensuring compatibility with other taxonomic data systems. Further, BCO outlines how specimens are preserved and curated, ensuring long-term data continuity and access. Its primary goal is to create a uniform structure for biological collections data, facilitating the integration of datasets from multiple institutions, ultimately promoting more efficient data exchange and reuse in biodiversity research.

The Taxonomic Rank Ontology has been structured to represent the hierarchical system of biological classification, ranging from broad categories such as kingdom to specific ranks like species. The ontology models each taxonomic level as a subclass of the parent class taxonomic_rank, reflecting the natural progression of biological categorization. The main taxonomic ranks include kingdom, phylum, class, order, family, genus, and species, ensuring that the system adheres to standard taxonomy conventions. Additionally, specialized and infraranks such as infrakingdom, infraorder, cultivar, and subspecies have been incorporated to offer greater granularity in classifications, particularly in botany. This structured hierarchy enhances the ontology's ability to support detailed representation and analysis of biodiversity, facilitating research in species relationships and taxonomic studies.

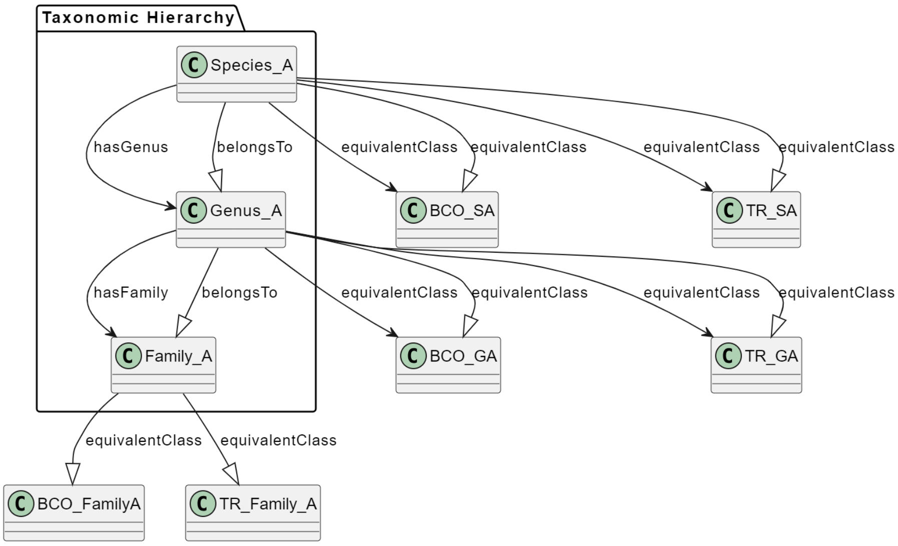


Figure 5.5: Taxonomic Hierarchy Diagram

5.3.1 Ontology Alignment Proces

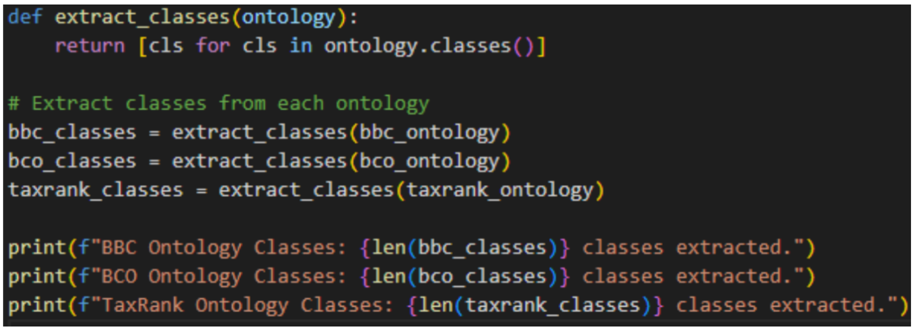
The following steps outline the process of aligning the BBC Ontology with BCO and ENVO. Each step involves extracting classes, matching them using string similarity, aligning taxonomic and environmental data, and saving the aligned ontology.

1. **Loading and Parsing Ontology Files**
   The first step in ontology alignment is the creation of RDF (Resource Description Framework) graphs, which represent the relationships between entities in a structured format. RDF is a widely-used data model that allows for the organization and integration of information, particularly for semantic data and web applications. The RDF model describes data as triples, consisting of a subject, predicate, and object, allowing for clear representation of relationships between resources. For example, in a butterfly ontology, a triple states that a species belongs to a specific genus: (Species A) — [hasGenus] —> (Genus B). This structure enables complex relationships to be modeled and integrated

across different datasets.
RDF triples are foundational for ontology alignment because they capture not only the entities in an ontology but also the relationships between them. The subject represents the entity being described, the predicate defines the property or relationship, and the object can either be another entity or a literal value such as a string or number. By building RDF graphs from these triples, we can represent intricate networks of relationships between taxonomic and biological collection data. In a biodiversity context, RDF graphs can illustrate how species relate to their habitats, taxonomic categories, or specimen data, forming the basis for alignment between different ontologies. Ontology alignment aims to link equivalent concepts across multiple datasets to ensure consistent representation and interpretation of data. RDF graphs are instrumental in achieving this alignment. When aligning the BBC Ontology with the Biological Collections Ontology (BCO) and the TaxRank Ontology, RDF triples serve as a mechanism to map similar or identical concepts. For instance, RDF can represent a butterfly species from the BBC Ontology as semantically equivalent to the same species in BCO. By creating these mappings, the alignment process ensures semantic consistency across datasets, allowing for integrated querying and data reuse.
Equivalence in ontology alignment is established through RDF triples by linking classes and properties that share the same meaning across ontologies. For example, a triple might represent the alignment of a Species class from the BBC Ontology with the corresponding Species class in BCO. The predicate in this case could be owl:equivalentClass, indicating that these two classes are considered semantically identical. This type of mapping not only ensures consistency in class definitions but also allows reasoning engines and queries to recognize these entities as the same, regardless of the ontology in which they originated.
RDF graphs also support the alignment of properties and relationships, extending the ontology integration beyond just classes. For instance, properties like hasGenus or hasSpecies in BBC can be aligned with similar properties in BCO and TaxRank. Using RDF to represent these connections facilitates the creation of a unified ontology that allows researchers to conduct comprehensive biodiversity analyses. By leveraging RDF graphs, ontologies like BBC, BCO, and TaxRank can be linked in a way that preserves semantic meaning while supporting advanced reasoning and interoperability across diverse datasets.

2. **Extracting Classes from the Ontologies**
In ontology alignment, extracting classes from RDF graphs is crucial, as these classes represent the primary concepts within a domain. Extracting and aligning the core taxonomic and biological collection classes ensures that key entities can be accurately represented and connected, which is essential for building a consistent and interoperable system across diverse ontologies. Taxonomic classes like Species and Genus are vital for biological ontologies because they provide the hierarchical framework for categorizing organisms. By defining how species relate to genera and other taxonomic ranks, these classes help maintain scientific consistency when describing biological entities. Likewise, classes like Sample and MaterialSample in BCO are essential for representing biological specimens collected during research. Aligning these taxonomic and biological

collection classes across ontologies like BBC, BCO, and TaxRank ensures that data from different sources can be seamlessly integrated, allowing researchers to make cross-dataset queries and conduct comprehensive analyses. The alignment process begins with classes because they define the ontology's fundamental structure, which all properties and relationships are built upon. Once the classes are aligned, the relationships between them,such as a species' connection to its genus or a sample's link to an organism can also be mapped. This approach allows individuals (instances of these classes) to be consistently aligned across datasets. By focusing on class alignment first, the ontology alignment process establishes a solid foundation that facilitates the mapping of properties, relationships, and individuals, ensuring interoperability and data consistency across diverse datasets.



```
def extract_classes(ontology):
    return [cls for cls in ontology.classes()]

# Extract classes from each ontology
bbc_classes = extract_classes(bbc_ontology)
bco_classes = extract_classes(bco_ontology)
taxrank_classes = extract_classes(taxrank_ontology)

print(f"BBC Ontology Classes: {len(bbc_classes)} classes extracted.")
print(f"BCO Ontology Classes: {len(bco_classes)} classes extracted.")
print(f"TaxRank Ontology Classes: {len(taxrank_classes)} classes extracted.")
```

3. **String-Based and Semantic Similarity Matching for Class and Property Alignment**

   In ontology alignment, identifying equivalent classes and properties across different ontologies is critical for ensuring consistent data representation and interpretation. This process involves determining whether classes like "Species" in the BBC Ontology correspond to the same class in other ontologies, such as the Biological Collections Ontology (BCO) or TaxRank. Both string-based similarity and semantic similarity techniques are employed to identify potential equivalence between classes. These methods help reconcile different naming conventions and terminological variations across ontologies, thus facilitating effective data integration.

   String-based similarity matching involves directly comparing the names of classes from different ontologies to assess how closely they align. For example, an ontology might use "Species_A" while another uses "BCO_SpeciesA" to refer to the same concept. Using algorithms SequenceMatcher from the difflib module, a similarity score ranging from 0.0 (no similarity) to 1.0 (perfect match) is computed to evaluate the match between two strings. A threshold, typically set around 0.7, helps determine when two classes are likely equivalent. This ensures that minor naming variations are captured, reducing the risk of false positives or negatives in the alignment process.

   While string-based matching focuses on the syntactical alignment of class names, semantic similarity goes deeper by comparing the underlying meanings of those names. This approach is especially useful when different ontologies use distinct terms for the same concept. Tools like WordNet — a lexical database of English

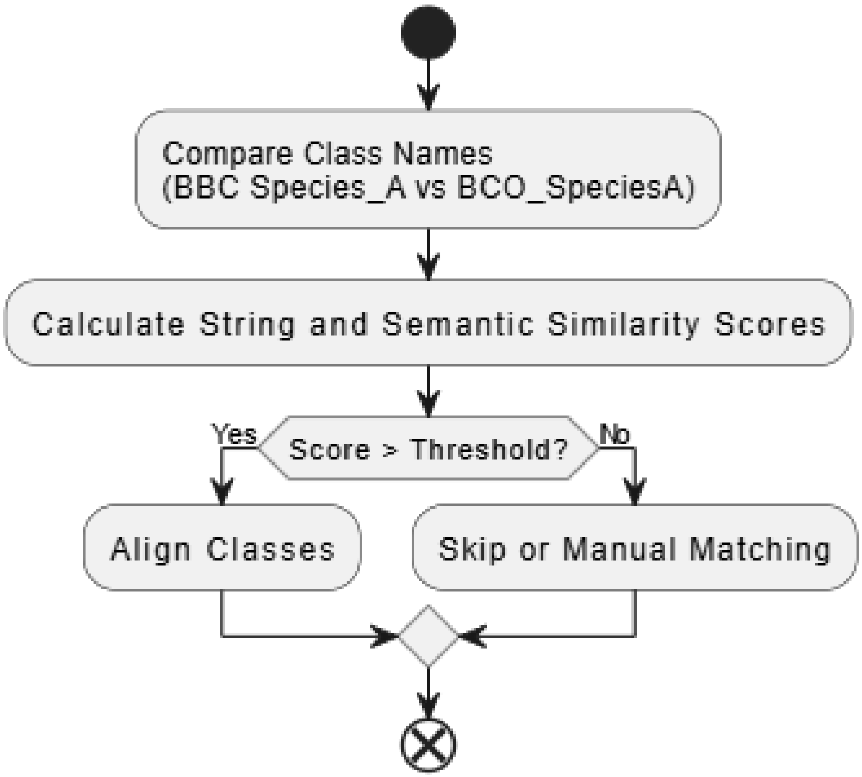
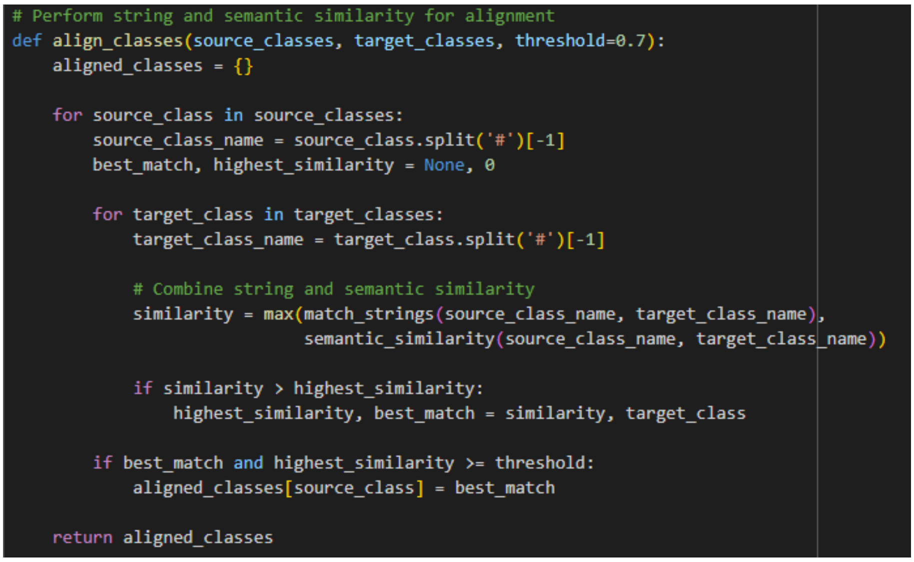

Compare Class Names
(BBC Species_A vs BCO_SpeciesA)
Calculate String and Semantic Similarity Scores
Score > Threshold?
Yes
No
Align Classes
Skip or Manual Matching

Figure 5.6: Similarity Matching Process

— allow for the calculation of semantic similarity by identifying the relationships between the words' synsets (sets of synonyms). Using algorithms such as Wu-Palmer similarity, the distance between meanings of class names is measured. Even if two class names do not match perfectly in terms of string similarity, their conceptual alignment can still be established if they share similar meanings in their synsets.

Combining string-based and semantic similarity enhances the robustness of the ontology alignment process. String similarity captures small variations in naming conventions, while semantic similarity addresses conceptual differences that string matching alone cannot detect. For example, a high string similarity score between "Species_A" and "BCO_SpeciesA," combined with a strong semantic match in WordNet synsets, ensures that these classes are recognized as equivalent. This comprehensive approach allows for more accurate alignment of ontologies within BBC, BCO, and TaxRank.

4. **Mapping Aligned Classes and Properties**
   Once matching classes and properties are identified during the ontology alignment process, the next step is to map these aligned entities to ensure they are recognized as semantically equivalent across different ontologies. This mapping involves linking taxonomic entities (e.g., species, genus) and specific instances (e.g., butterfly specimens) to their corresponding concepts in other ontologies, such as the Biological Collections Ontology (BCO) and TaxRank. Establishing these mappings is essential for enabling consistent, unified querying and analy-



```python
# Perform string and semantic similarity for alignment
def align_classes(source_classes, target_classes, threshold=0.7):
    aligned_classes = {}

    for source_class in source_classes:
        source_class_name = source_class.split('#')[-1]
        best_match, highest_similarity = None, 0

        for target_class in target_classes:
            target_class_name = target_class.split('#')[-1]

            # Combine string and semantic similarity
            similarity = max(match_strings(source_class_name, target_class_name),
                             semantic_similarity(source_class_name, target_class_name))

            if similarity > highest_similarity:
                highest_similarity, best_match = similarity, target_class

        if best_match and highest_similarity >= threshold:
            aligned_classes[source_class] = best_match

    return aligned_classes
```

sis across different datasets.

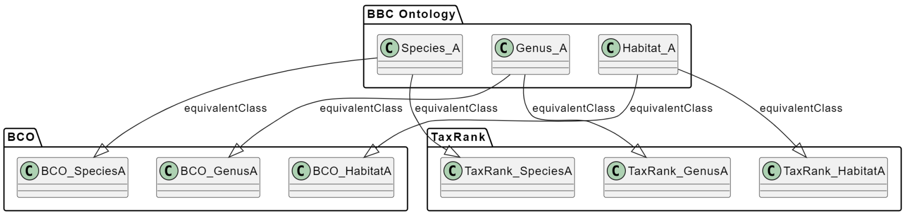
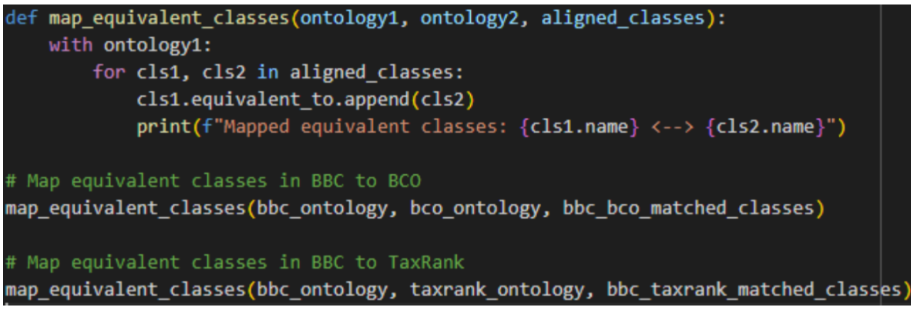


Figure 5.7: Class Matching Diagram

To formalize the equivalence of the entities, the OWL properties OWL.equivalentClass and OWL.equivalentProperty is used. OWL.equivalentClass is used to declare that two classes from different ontologies represent the same concept even if the class name is different, while OWL.equivalentProperty indicates that two properties have the same meaning and can be used interchangeably. By applying these properties, reasoning engines and query systems can treat the linked classes and properties as identical, regardless of differences in their naming or structure across ontologies. The mapping process starts by identifying equivalent classes and properties for taxonomic entities. Once these equivalences are established, they are mapped into an RDF graph using the OWL properties mentioned earlier. If the class BBC is found to be equivalent to BCO, this relationship is formalized in the RDF graph using OWL.equivalentClass. Similarly, properties like BBC and BCO are linked using OWL.equivalentProperty to ensure they are treated as equivalent across the ontologies. By mapping aligned classes and properties in this way, the RDF graph creates a formalized representation of equivalence. This ensures that the same concepts and relationships are recognized across all involved ontologies.

The above steps can be simplified and shown in the form of a flowchart. Figure 5.8 illustrates this flowchart briefing all the steps in the ontology alignment process.



```
def map_equivalent_classes(ontology1, ontology2, aligned_classes):
    with ontology1:
        for cls1, cls2 in aligned_classes:
            cls1.equivalent_to.append(cls2)
            print(f"Mapped equivalent classes: {cls1.name} <--> {cls2.name}")

# Map equivalent classes in BBC to BCO
map_equivalent_classes(bbc_ontology, bco_ontology, bbc_bco_matched_classes)

# Map equivalent classes in BBC to TaxRank
map_equivalent_classes(bbc_ontology, taxrank_ontology, bbc_taxrank_matched_classes)
```

5.3.2 Classes Used in the Ontology Alignment Process

The integration of ontologies for the Barcant Butterfly Collection involves aligning classes from the BBC Ontology (Butterfly Biodiversity and Observation Ontology), BCO (Biological Collections Ontology), and TaxRank Ontology (Taxonomic Rank Ontology). These ontologies each contribute critical classes that serve distinct roles in capturing taxonomic, observational, and specimen data, enabling comprehensive and detailed biodiversity research. Below is a detailed breakdown of the classes used from each ontology and their role in this integration.

1. **Classes from the BBC Ontology**
   The BBC Ontology focuses on modeling taxonomic and observational data specifically related to butterflies. Key classes from this ontology include TaxonKey and SpeciesKey, which serve as unique identifiers for taxa (genus or species) and species, respectively. The ScientificName class represents the official scientific name of butterfly species, ensuring taxonomic accuracy. Observation-related details are captured through classes like EventDate, which records the date of the butterfly observation, and GeographicalLocation, which specifies where the observation took place. Additionally, RecordedBy and IdentifiedBy attribute the observation and identification of the butterfly to the responsible person or entity. The Observation class encapsulates the entire observational event, including details such as butterfly behavior and environmental factors. These classes establish the backbone for linking taxonomic data with specific observational metadata.

2. **Classes from the Biological Collections Ontology (BCO)**
   The BCO Ontology provides a framework for capturing information about biological specimen collections and associated events. The Sample and MaterialSample classes refer to the biological specimen collected, specifying the type of sample, such as preserved, fossilized, or living. The Event class represents the occurrence during which the specimen was collected, while EventAttribute records specific attributes of the event, such as environmental conditions. The Location and GeologicalContext classes provide geographical and environmental context, detailing where the collection occurred and describing factors. Occurrence and Organism represent the occurrence of the specimen and describe the organism from which the sample was collected, respectively. These classes are crucial for linking specimen data with environmental metadata and event details.

3. **Classes from the TaxRank Ontology**

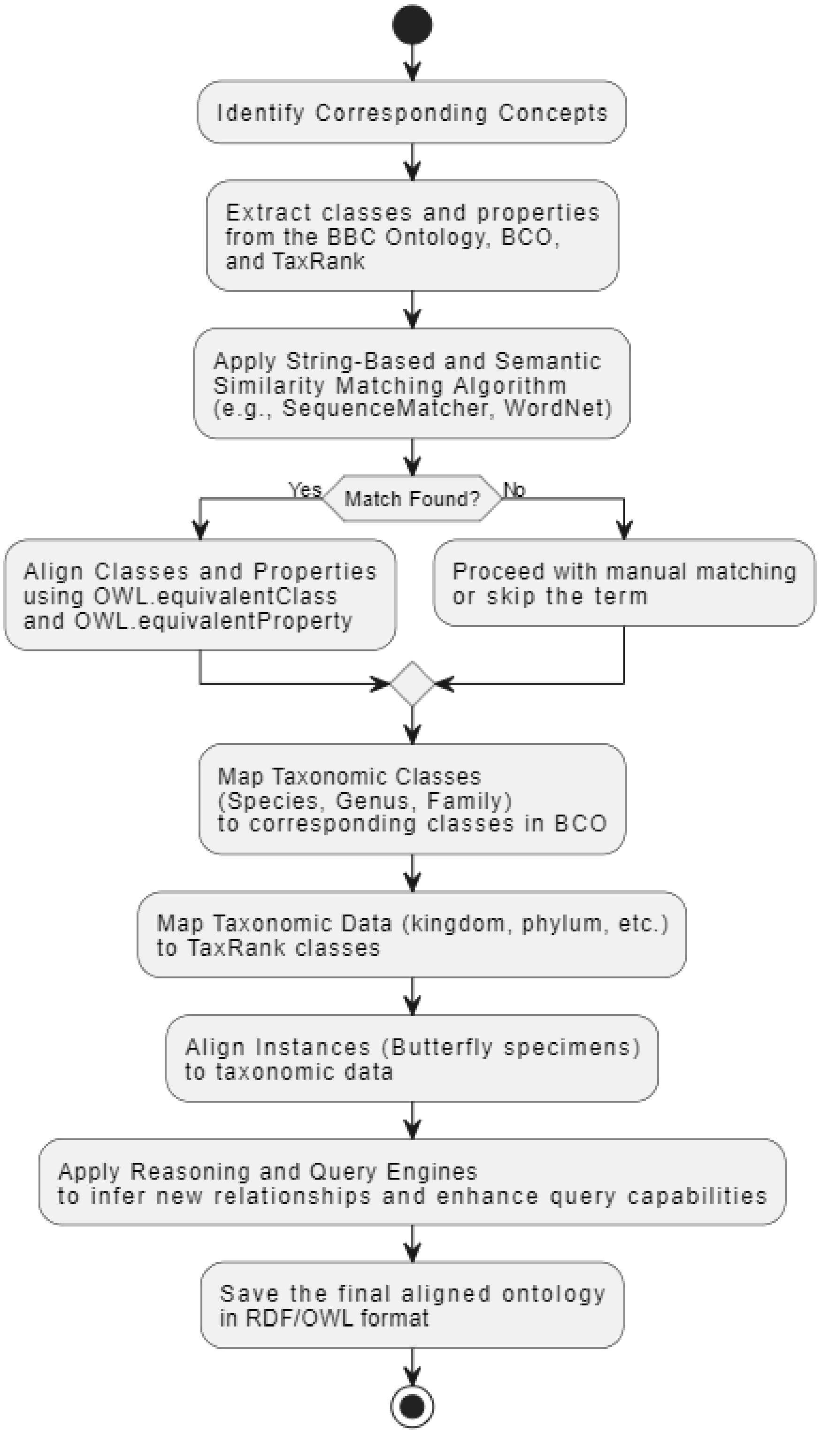

Identify Corresponding Concepts
Extract classes and properties from the BBC Ontology, BCO, and TaxRank
Apply String-Based and Semantic Similarity Matching Algorithm (e.g., SequenceMatcher, WordNet)
Yes
Match Found?
No
Align Classes and Properties using OWL.equivalentClass and OWL.equivalentProperty
Proceed with manual matching or skip the term
Map Taxonomic Classes (Species, Genus, Family) to corresponding classes in BCO
Map Taxonomic Data (kingdom, phylum, etc.) to TaxRank classes
Align Instances (Butterfly specimens) to taxonomic data
Apply Reasoning and Query Engines to infer new relationships and enhance query capabilities
Save the final aligned ontology in RDF/OWL format

Figure 5.8: Flowchart of ontology alignment process

The TaxRank Ontology plays a central role in organizing taxonomic information by capturing the hierarchical structure of biological classification. The Taxonomic_Rank class serves as the top-level concept, encompassing various taxonomic levels such as Kingdom, Phylum, Class, Order, Family, Genus, and Species. These ranks define the relationships between different levels of biological classification, ensuring a consistent and structured representation of species data. Further granularity is achieved with classes like Subspecies, Infraspecies, and Varieties, which represent finer taxonomic divisions. These taxonomic classes are foundational to ensuring that species are correctly organized and classified within the ontology, aligning closely with the biological classification system.

5.4 Generating knowledge graphs

Generating a knowledge graph from an OWL (Web Ontology Language) ontology provides a structured way to visualize relationships between entities in a dataset. OWL ontologies encode hierarchical and semantic links, and Python libraries such as ‘owlready2‘ and ‘networkx‘ allow for efficient construction and visualization of these relationships as directed graphs. This method enables researchers to interpret the ontology’s structure by transforming abstract classes and relationships into an intuitive graphical format.

Using ‘owlready2‘, the ontology is loaded and parsed, extracting entities (nodes) and their relationships (edges), such as subclass hierarchies. With ‘networkx‘, these entities and relationships are represented as a directed graph, where each class becomes a node and subclass relationships form the edges. Visualization through ‘matplotlib‘ helps to clearly depict the hierarchical and relational structure of the ontology, making the data easier to explore and analyze.

This approach allows for a clearer understanding of complex datasets, especially in fields like biodiversity, where taxonomy and hierarchical relationships are key. Visualizing the ontology as a graph highlights critical relationships, enabling researchers to quickly identify patterns and connections.

In summary, generating knowledge graphs from OWL ontologies offers an effective means of exploring complex data structures. This technique provides a clear, interactive view of the underlying relationships, making it a valuable tool for both data analysis and decision-making in research contexts.

Figure 5.9 presents a densely interconnected knowledge graph generated from the ontology, with nodes representing entities like "species," "sample," and "observation," and edges denoting relationships among them. The hierarchical structure, depicted in a circular layout, positions general concepts at the center, with specific ones radiating outward. The identifiers such as "TAXARANK" and "BCO" suggest a focus on biodiversity and taxonomy, aiding researchers in exploring classifications and connections within the ontology. By integrating reasoning via an OWL reasoner like Pellet through the owlready2 library, additional relationships such as subclasses and equivalences can be inferred based on logical axioms. This reasoning enriches the knowledge graph by revealing hidden connections, ensuring consistency, and enhancing query capabilities.

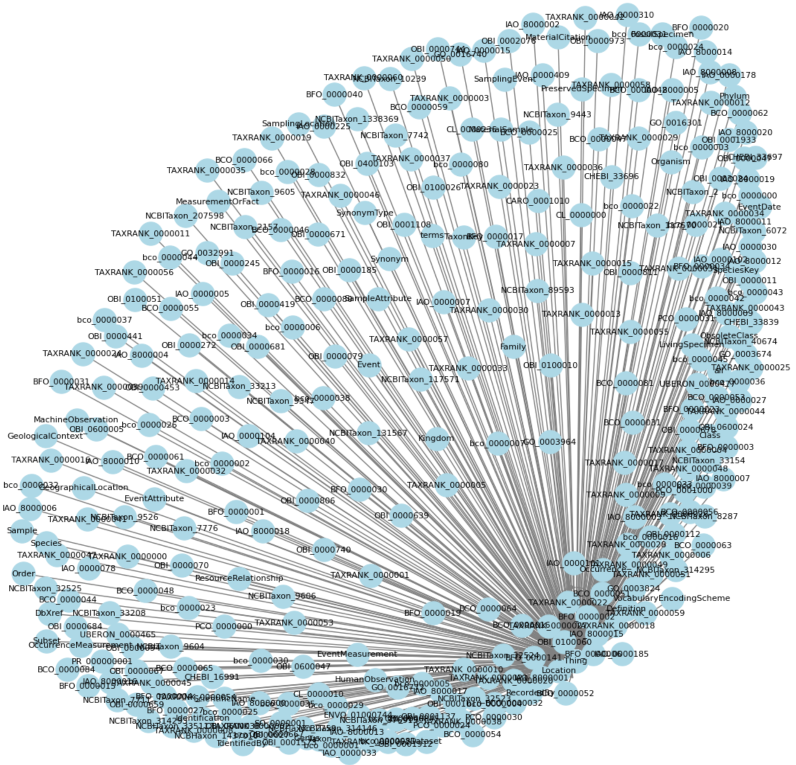


Figure 5.9: The resulting knowledge graph

Chapter 6

Applications

In this chapter, we explore the practical applications of the methodologies and techniques developed throughout this research. By applying these methods in real-world contexts, we aim to demonstrate the utility and relevance of our approach in addressing challenges within the domain. This application not only validates the theoretical concepts but also provides insights into their effectiveness and adaptability across various scenarios. Through this chapter, we illustrate how the research outcomes can be translated into actionable solutions, offering potential benefits for practitioners and stakeholders in the field.

6.1 Ecological Range Mapping from BBC Ontology

Ecological range maps offer a data-driven approach to visualizing species distributions by combining species occurrence data with environmental variables such as climate, habitat type, and ecological interactions. These maps capture the dynamic nature of species distributions in response to changing environmental conditions. The maps are generated using real-time data and computational models, providing a scalable solution for biodiversity research and conservation.

6.1.1 Strengths of Ecological Range Maps

Ecological range maps offer a data-driven approach to visualizing species distributions by combining species occurrence data with environmental variables such as climate, habitat type, and ecological interactions. These maps are particularly valuable in capturing the dynamic and complex nature of species distributions in response to changing environmental conditions. They are generated through the integration of real-time data and computational models, providing a flexible and scalable solution for biodiversity research and conservation.

One of the key strengths of ecological range maps lies in their ability to incorporate large datasets from diverse sources, such as species occurrence records, climate data, and habitat information. This data-driven approach enables the creation of maps that reflect the most up-to-date and comprehensive information available. Unlike expert range maps, which rely heavily on expert interpretation, ecological range maps are grounded in empirical data, reducing the potential for bias and subjective errors. This leads to a more objective representation of species distributions that can be continuously updated as new data becomes available. Ecological range maps are inherently adaptable, making them better suited to reflect the dynamic nature of ecosystems. As environmental conditions change, these maps can be quickly updated to account for new data, ensuring that they remain relevant and accurate. This adaptability is particularly important in the context of climate change, where species distributions are rapidly shifting. By using ecological range maps, researchers and conservationists can more effectively monitor and respond to these changes, ensuring that conservation strategies are based on the most current information.The computational methods used in ecological range mapping, such as species distribution models, allow for scalability across different spatial and temporal scales. This makes ecological range maps highly versatile, applicable to both local and global studies. The ability to model potential distributions based on various environmental scenarios also provides valuable insights into future species distributions under different climate change models. This scalability and broad applicability make ecological range maps a powerful tool for addressing complex ecological questions and informing conservation efforts on multiple levels.

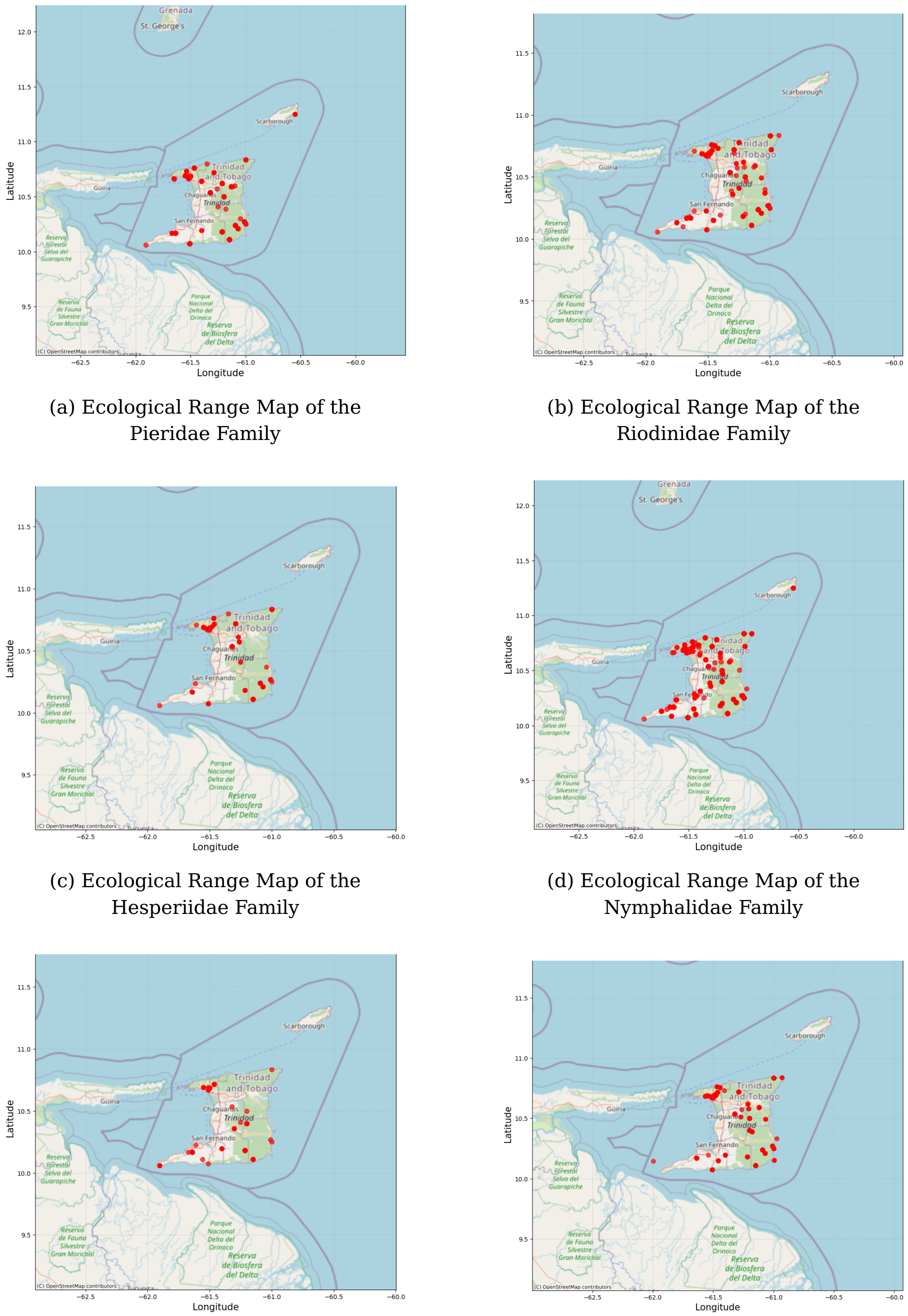
Ecological range maps benefit from the continuous integration of new data sources, including citizen science contributions, remote sensing data, and advances in ecological modeling techniques. This constant influx of data enables the maps to evolve over time, improving their accuracy and predictive power. In contrast, expert range maps may become outdated as they rely on static knowledge bases that may not be as frequently updated. The ability of ecological range maps to incorporate new information in real-time makes them more relevant in rapidly changing environments.

In the Barcant Butterfly Collection the ontology has classes and properties to capture data about species locations, such as GeographicLocation, DecimalLatitude, DecimalLongitude and EventDate. Next, maps are generated to display butterfly occurrences using data from the ontologies, stored in RDF format. First, load the RDF graph from the created ontologies, extract the names of various species and their locations, and then organize this information into a GeoDataFrame using GeoPandas. Assign the GeoDataFrame a Coordinate Reference System (CRS) of WGS84 to align with global positioning standards. Next, plot the data on a map, using Matplotlib to display red dots at each butterfly sighting location. A background map from OpenStreetMap is added using Contextily to provide geographical context. The view is zoomed in on Trinidad and Tobago to enhance readability. These maps illustrate the distribution of various butterfly species.

In the course of generating biodiversity maps from the Barcant Butterfly Collection RDF dataset, we initially anticipated producing six maps, each representing a unique family of butterfly species. However, upon execution of the mapping algorithm, ten maps were generated, indicating a higher number of unique families than expected. This discrepancy required further investigation into the dataset's structure.

To understand the distribution of species across families, the dataset was examined by grouping species according to their family names and counting the species associated with each family. The families obtained were Pieridae, Lycaenidae, Hesperiidae, Nymphalidae, Sesiidae, Papilionidae, Riodinidae, Ephydridae, Noctuidae, Zygaenidae. Since the objective was to generate six numbers of maps with main families, the code was adjusted accordingly.

The unexpected generation of additional maps highlights the importance of thor-


Grenada
St. George's
Scarborough
Trinidad and Tobago
Chaguanas
Trinidad
San Fernando
Güiria
Reserva Forestal Selva del Guarapiche
Reserva de Fauna Silvestre Gran Morichal
Parque Nacional Delta del Orinoco
Reserva de Biosfera del Delta
Tucupita
(C) OpenStreetMap contributors
Latitude
Longitude
12.0
11.5
11.0
10.5
10.0
9.5
−62.5
−62.0
−61.5
−61.0
−60.5
−60.0

(a) Ecological Range Map of the Pieridae Family

(b) Ecological Range Map of the Riodinidae Family

(c) Ecological Range Map of the Hesperiidae Family

(d) Ecological Range Map of the Nymphalidae Family

(e) Ecological Range Map of the Papilionidae Family

(f) Ecological Range Map of the Lycaenidae Family

Figure 6.1: Ecological Range Maps of the Butterfly Occurrences in Various Families

ough data validation in biodiversity research. The presence of more families than expected could indicate a inconsistent dataset, it underscores the potential for errors due to inconsistent taxonomy. By investigating the family distribution and standardizing family names, we can ensure accurate and consistent visualizations of species distribution.

6.2 Comparing BBC with WDPA Database

The Protected Planet (WDPA) dataset provides detailed information about protected areas around the world. It includes metadata on each protected area, such as its name, designation, IUCN category, and geographical location. Specifically, the dataset focuses on conservation areas like nature reserves and national parks, categorizing them by their protection status and management authority. In the case of the Trinidad and Tobago (TTO) dataset, it contains information on protected areas like the "Rochard Douglas Reserve" and the "Tacarigua Reserve," providing details such as the designation type (e.g., nature reserve, national park), IUCN category, and other administrative data like ownership and management.(Obtained from [66])

The comparison of species from the Barcant Butterfly Collection with those found within specific protected areas in the Trinidad and Tobago (TTO) region offers critical insights into the biodiversity preserved within these zones. By analyzing the geographic coordinates and the distribution of butterfly species across various protected areas, we can gauge how effectively these zones are maintaining species diversity.The comparison reveals which protected areas host the most diverse populations of butterflies, showing areas of high conservation success. It also allows for identifying gaps where certain species may be underrepresented or where protected zones are not fully safeguarding the full range of species documented in Barcant's collection. This can be vital for refining conservation strategies and prioritizing areas that require additional protection or resource allocation.

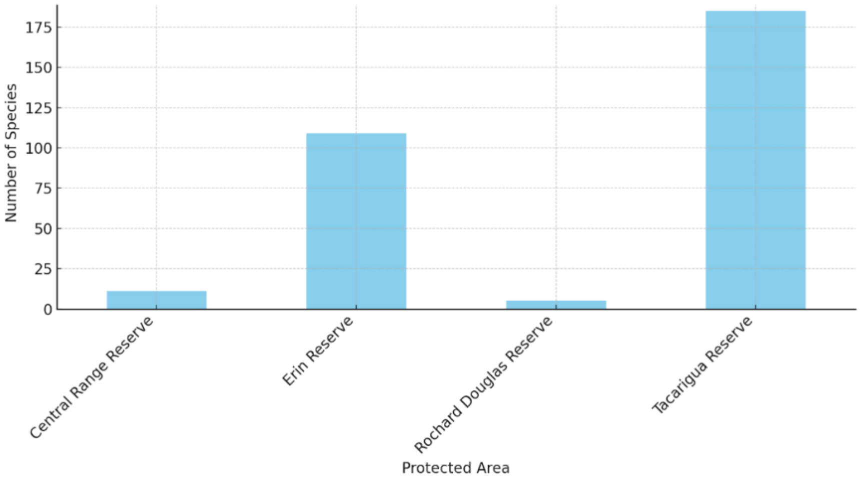


Figure 6.2: Number of Unique Species in Each Protected Area

The bar chart illustrates the distribution of butterfly species across four protected areas: Tacarigua Reserve, Erin Reserve, Central Range Reserve, and Rochard Douglas Reserve. The Tacarigua Reserve holds the largest number of species, with nearly 250 species recorded, followed by the Erin Reserve with approximately 150 species. Both Central Range Reserve and Rochard Douglas Reserve show significantly fewer species, with under 50 species each. This visualization highlights the biodiversity richness in Tacarigua and Erin Reserves compared to the other protected areas, emphasizing their importance in conservation efforts.

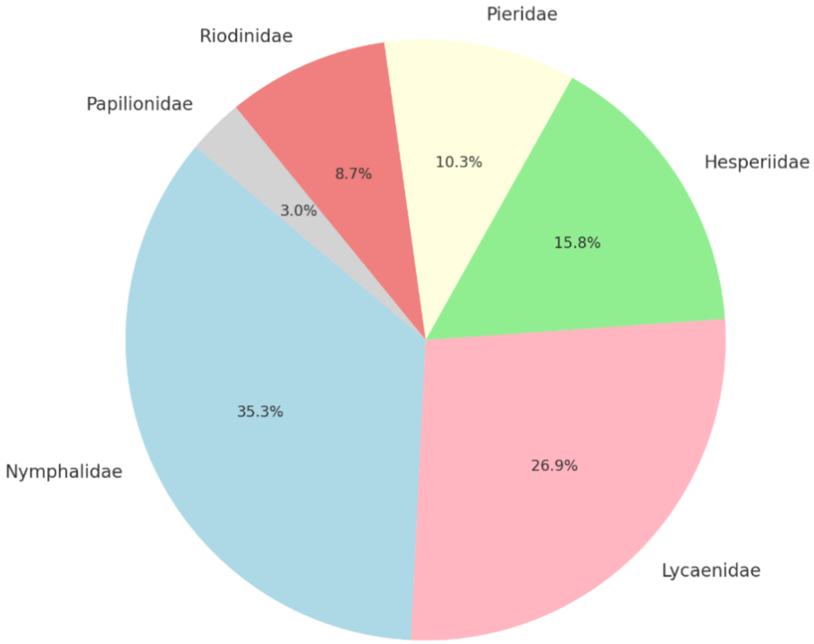


Figure 6.3: Species Distribution by Family in Protected Areas

The pie chart displays the representation of the butterfly families within the protected areas. The Nymphalidae family dominates, accounting for 35.3% of the total species, followed by Lycaenidae with 26.9%, and Hesperiidae with 15.8%. Pieridae represents 10.3% of the species, while Riodinidae and Papilionidae are less prominent, contributing 8.7% and 3.0%, respectively. This breakdown highlights the significant diversity within the Nymphalidae and Lycaenidae families in comparison to the other families, suggesting their strong presence in the protected areas covered by the dataset.

The scatter plot displays the geographic distribution of butterfly species based on their latitude and longitude. The species are spread across a range of latitudes from around 10.1 to 10.7 and longitudes from approximately -61.7 to -61.2. This visualization shows clusters of species in certain areas, indicating regions with higher species density. The spread of species across these coordinates reflect variations in habitat suitability or protection efforts within these geographical zones. Such a plot is useful for identifying biodiversity hotspots and understanding the spatial patterns of species distribution in relation to environmental factors or conservation areas.

Given that Trinidad and Tobago is home to unique ecosystems and rich biodiversity, understanding how well its protected areas conserve butterfly species can be a benchmark for broader conservation efforts in the region. The findings from this comparison may help policymakers and conservationists adjust management practices to

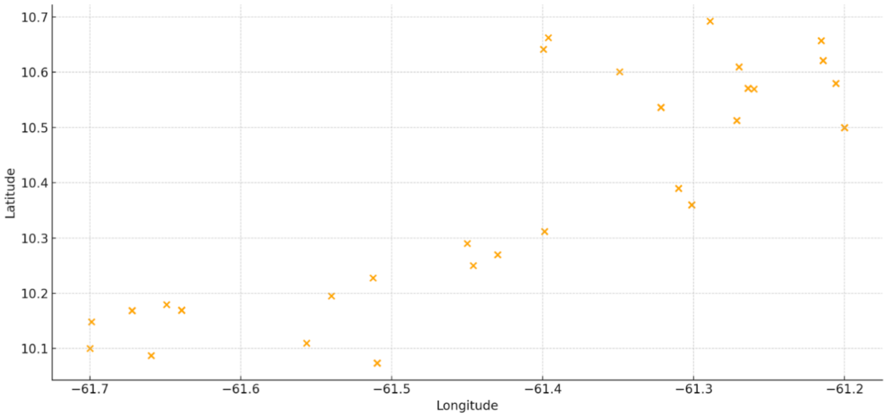


Figure 6.4: Species Count by Geographic Latitude and Longitude

improve butterfly preservation, contribute to long-term biodiversity goals, and ensure that the rich natural heritage of TTO is sustained. Furthermore, these insights can be used to monitor changes in butterfly populations over time, helping to track the impacts of environmental changes, human activities, and climate on species distribution. This holistic understanding of species conservation in the TTO region not only benefits butterflies but also plays a crucial role in the broader ecosystem health of the islands.

6.3 Evaluation of the Aligned Ontology

A robust and thorough evaluation of an ontology is essential to ensure its quality, usability, and adherence to the domain it seeks to model. To assess the developed ontology, a task-based evaluation approach was employed, leveraging SPARQL queries to inspect critical components such as class definitions, instance data, property relationships, and overall structural consistency. This section outlines the key evaluation tasks, discussing both the methodology and results obtained during the process.

6.3.1 SPARQL Query for Retrieving Class Details from Ontology

The provided SPARQL query is designed to retrieve unique classes, along with their labels and comments, from an ontology using RDF (Resource Description Framework) and OWL (Web Ontology Language) prefixes. The query focuses on selecting distinct classes defined within the ontology (identified by the '?class' variable) by specifying that each '?class' is of type 'owl:Class'. It then attempts to retrieve associated labels and comments for these classes, if available, by using the 'OPTIONAL' keyword to include 'rdfs:label' and 'rdfs:comment' properties. This query can be particularly useful for ontology documentation, as it allows users to get descriptive information about each class, making it easier to understand the ontology structure and the roles of various classes without needing to delve into each class individually. The use of 'DISTINCT' ensures that each class is listed only once, avoiding redundancy in the output.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX ex: <http://example.org/merged_ontology.owl#>

SELECT DISTINCT ?class ?label ?comment
WHERE {
    ?class a owl:Class .
    OPTIONAL { ?class rdfs:label ?label }
    OPTIONAL { ?class rdfs:comment ?comment }
}
```

6.3.2 SPARQL Query for Retrieving Instances and Their Classes from Ontology

This SPARQL query is designed to extract unique instances and their corresponding classes from an RDF-based ontology. By selecting distinct pairs of '?instance' and '?class', it identifies each instance in the ontology along with its associated class type, defined through the 'rdf:type' relationship. The '?instance' variable represents individual instances within the ontology, while '?class' represents the class that each instance belongs to. This query is useful for obtaining a structured view of the data within an ontology, as it outlines the types of instances present and their classifications, providing insight into the data model and instance distributions across classes. The use of 'DISTINCT' ensures that each instance-class pair is listed only once in the results.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?instance ?class
WHERE {
    ?instance rdf:type ?class .
}
```

6.3.3 SPARQL Query for Retrieving All Triples in an RDF Dataset

This SPARQL query is structured to retrieve all unique triples from an RDF (Resource Description Framework) dataset. Each triple consists of a '?subject', '?property', and '?object', representing the core components of RDF statements where '?subject' is the resource being described, '?property' is the predicate or attribute of the subject, and '?object' is the value associated with that property. By selecting distinct combinations of these variables, the query provides a comprehensive view of all relationships and data points within the RDF dataset. This approach is valuable for obtaining a complete snapshot of the data structure, enabling users to analyze connections and understand the overall dataset organization. The 'DISTINCT' keyword ensures that each triple is returned only once, preventing duplicate entries in the output.

```
SELECT DISTINCT ?subject ?property ?object
WHERE {
    ?subject ?property ?object .
}
```

6.3.4 SPARQL Query for Retrieving Subclass and Superclass Relationships in an Ontology

This SPARQL query is designed to retrieve subclass-superclass relationships within an ontology. Using the ‘rdfs:subClassOf‘ property, it identifies each ‘?subClass‘ and its corresponding ‘?superClass‘. The ‘?subClass‘ variable represents a more specific class within the ontology, while ‘?superClass‘ denotes a more general class that the subclass inherits from or extends. This hierarchical structure, commonly found in ontologies, helps organize concepts from general to specific. This query is especially useful for understanding the class hierarchy and for applications that require traversing these relationships, such as reasoning engines or ontology visualization tools.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?subClass ?superClass
WHERE {
    ?subClass rdfs:subClassOf ?superClass .
}
```

6.3.5 SPARQL Query for Retrieving Individuals Not Classified as Classes

This SPARQL query aims to identify unique individuals in an RDF dataset that are not themselves classified as classes. The ‘?individual‘ variable represents entities with a specified type (‘rdf:type ?type‘), but the query applies a filter to exclude any individuals whose type is defined as an ‘rdfs:Class‘. This distinction is made using the ‘FILTER NOT EXISTS‘ clause, which removes any entries where ‘?type‘ has an ‘rdf:type‘ of ‘rdfs:Class‘. This query is particularly useful in cases where we need to focus on instances or objects in the dataset without including ontology structures or classes, effectively narrowing down the output to concrete individual data points.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?individual
WHERE {
    ?individual rdf:type ?type .
    FILTER NOT EXISTS { ?type rdf:type rdfs:Class }
}
```

The results from each of these SPARQL queries were extracted and saved in CSV format to facilitate ease of access and further analysis. These CSV files, containing comprehensive data on class hierarchies, instance classifications, and subclass-superclass relationships, were subsequently uploaded to a GitHub repository alongside the SPARQL query codes and additional documentation. This repository serves as a consolidated resource for all relevant data and code utilized in this research. Interested readers and researchers can access these files directly at the following link:

https://github.com/raksh07/Thesis. By providing open access to these materials, we aim to support reproducibility and enable further exploration and validation of our findings.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This study explores the potential of ontology matching and knowledge graph techniques to improve the analysis and utility of biodiversity data from historical collections, with a specific focus on the Barcant Butterfly Collection. In response to RQ1, the Barcant Butterfly Collection ontology was designed to incorporate essential ontological concepts and relationships needed to capture the nuanced details of butterfly biodiversity. Key concepts in this ontology include taxonomic hierarchy, which encompasses classification levels from family to species; geographical distribution, mapping butterfly population locations; and ecological interactions, connecting butterflies with habitats and associated species. Together, these concepts ensure a robust framework capable of representing both scientific and ecological complexities within the butterfly data.

Addressing RQ1.a, the study carefully selected standardized vocabularies and terminologies from established ontologies, notably Darwin Core (DwC) and the Biological Collections Ontology (BCO). These vocabularies provided a set of terms that could be seamlessly integrated without redefinition, thereby minimizing redundancy and facilitating compatibility with broader biodiversity data platforms. These lexicons were particularly valuable as they encompass taxonomic and ecological terminology well-suited to butterfly collection data.

In response to RQ1.b, among the array of available biodiversity ontologies, Darwin Core and BCO emerged as the most compatible frameworks for the Barcant Butterfly Collection due to their established structure, widespread use, and suitability for historical taxonomic data. Additionally, the TaxRank ontology was selected for its emphasis on hierarchical taxonomic data, aligning closely with the detailed classification requirements of butterfly biodiversity. This combination of ontologies proved optimal for enhancing the relevance of the Barcant Collection and enabling its integration with modern biodiversity databases.

Through the alignment of these ontologies, the study produced knowledge graphs that uncovered previously hidden relationships, enabling sophisticated querying capabilities to explore species interactions, habitat overlaps, and broader biodiversity patterns. The generation of ecological range maps further enriched the analysis, providing powerful visualization tools for assessing species distribution, highlighting the collection's potential for identifying biodiversity hotspots, and underscoring its applicability to conservation efforts, including integration with the World Database on Protected Areas (WDPA).

The evaluation of the Barcant Butterfly Collection ontology addressed RQ2 through specific performance metrics and targeted queries, designed to assess taxonomic precision, efficacy in species distribution analysis, and relevance to conservation research. Queries examining species-habitat relationships, taxonomic hierarchies, and distribution-based hotspot identification demonstrated the ontology's practical utility and robustness. These evaluation results affirmed the ontology's value as a bridge between historical taxonomic records and contemporary biodiversity science, opening new avenues for conservation and scientific research.

7.2 Future Work

Several promising avenues for future research have been identified as a result of this study. First, while the ontology alignment in this work focused on taxonomic and geographic data, future efforts could extend this to include more complex ecological interactions, such as species competition, symbiosis, and predation. Incorporating these ecological dimensions into the ontology would enable a more comprehensive understanding of biodiversity dynamics.

Second, there is significant potential for enhancing the scalability of ontology alignment and knowledge graph generation by incorporating machine learning and artificial intelligence (AI) techniques. The use of automated tools such as machine learning classifiers and natural language processing (NLP) could streamline the ontology alignment process, improving both speed and accuracy when dealing with large, heterogeneous datasets. In particular, AI could assist in identifying complex semantic relationships between entities that are not immediately apparent through traditional matching techniques.

Additionally, this research could be extended by incorporating real-time data streams from citizen science platforms, environmental monitoring networks, and remote sensing technologies. By integrating dynamic data sources, the knowledge graphs could be continuously updated, allowing for real-time analysis of biodiversity changes and enabling more timely conservation interventions. This would also facilitate predictive modeling, enabling researchers to forecast potential changes in species distributions and ecosystem health in response to factors such as climate change and habitat loss.

Another potential extension is the application of this methodology to other historical and contemporary biological collections globally. The successful alignment and enrichment of the Barcant Butterfly Collection demonstrate that similar techniques could be applied to other underutilized biological datasets, contributing to the creation of a comprehensive, globally integrated biodiversity knowledge system. This would not only enhance scientific research but also support conservation initiatives by making biodiversity data more accessible, interoperable, and actionable.

Lastly, future work should focus on enhancing the visualization tools used for biodiversity analysis. While the current study utilized ecological range maps, more sophisticated visualization techniques such as interactive knowledge graphs, geospatial heatmaps, and 3D modeling could provide deeper insights and improve stakeholder engagement. Such tools could facilitate communication between researchers, conservationists, and policymakers, thereby enhancing the impact of biodiversity data on decision-making processes.

In conclusion, this research lays the groundwork for the continued development of ontology-driven biodiversity informatics, with significant implications for global biodiversity research and conservation. By refining and expanding the methods and technologies presented here, future studies can further bridge the gap between historical biological records and modern conservation strategies, ensuring the preservation of global biodiversity for generations to come.

Appendix: Repository of Code and Results

All files associated with this research, including SPARQL query codes, OWL files, CSV results, and additional data, are compiled in a GitHub repository. This repository contains the full range of outputs from the analysis, such as ontology files (in OWL format), result datasets in CSV, and all SPARQL queries used for data extraction and analysis. These resources are provided to facilitate reproducibility, allow for in-depth review, and support further research applications. The complete set of files is openly accessible at `https://github.com/raksh07/Thesis`, ensuring transparency and ease of access for any readers or researchers interested in extending or examining this work.

Bibliography

[1] Holly K Kindsvater, Nicholas K Dulvy, Cat Horswill, Maria-José Juan-Jordá, Marc Mangel, and Jason Matthiopoulos. Overcoming the data crisis in biodiversity conservation. *Trends in Ecology & Evolution*, 33(9):676–688, 2018.

[2] Robert P Guralnick, Andrew W Hill, and Meredith Lane. Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters*, 10(8):663–672, 2007.

[3] Christian König, Patrick Weigelt, Julian Schrader, Amanda Taylor, Jens Kattge, and Holger Kreft. Biodiversity data integration—the significance of data resolution and domain. *PLoS Biology*, 17(3):e3000183, 2019.

[4] Robert P Anderson, Miguel B Araújo, Antoine Guisan, Jorge M Lobo, Enrique Martínez-Meyer, A Townsend Peterson, and Jorge M Soberón. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Frontiers of Biogeography*, 12(3), 2020.

[5] Roderic DM Page. Dna barcoding and taxonomy: dark taxa and dark texts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702):20150334, 2016.

[6] Jarrett Blair, Rodger Gwiazdowski, Andrew Borrelli, Michelle Hotchkiss, Candace Park, Gleannan Perrett, and Robert Hanner. Towards a catalogue of biodiversity databases: An ontological case study. *Biodiversity Data Journal*, 8, 2020.

[7] Hemchandranauth Sambhu and Alliea Nankishore. Butterflies (lepidoptera) of guyana: A compilation of records. *Zootaxa*, 4371(1):1–187, 2018.

[8] Eric H Fegraus, Sandy Andelman, Matthew B Jones, and Mark Schildhauer. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (eml) and principles for metadata creation. *Bulletin of the Ecological Society of America*, 86(3):158–168, 2005.

[9] Sara E Miller, Lisa N Barrow, Sean M Ehlman, Jessica A Goodheart, Stephen E Greiman, Holly L Lutz, Tracy M Misiewicz, Stephanie M Smith, Milton Tan, Christopher J Thawley, et al. Building natural history collections for the twenty-first century and beyond. *BioScience*, 70(8):674–687, 2020.

[10] Thamara Zacca, Freddy Bravo, and Maíra Xavier Araújo. Butterflies (lepidoptera: Papilionoidea and hesperioidea) from serra da jibóia, bahia state, brazil. *EntomoBrasilis*, 4(3):139–143, 2011.

[11] Michael Iannacone, Shawn Bohn, Grant Nakamura, John Gerth, Kelly Huffer, Robert Bridges, Erik Ferragut, and John Goodall. Developing an ontology for cyber security knowledge graphs. In *Proceedings of the 10th Annual Cyber and Information Security Research Conference*, pages 1–4, 2015.

[12] Yves R Jean-Mary, E Patrick Shironoshita, and Mansur R Kabuka. Ontology matching with semantic verification. *Journal of Web Semantics*, 7(3):235–251, 2009.

[13] Karin Koogan Breitman, Marco Antonio Casanova, and Walter Truszkowski. Ontology in computer science. *Semantic Web: Concepts, Technologies and Applications*, pages 17–34, 2007.

[14] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? *Handbook on Ontologies*, pages 1–17, 2009.

[15] John Deck, Robert Guralnick, Ramona Walls, Stanley Blum, Melissa Haendel, Andréa Matsunaga, and John Wieczorek. Meeting report: identifying practical applications of ontologies for biodiversity informatics. *Standards in Genomic Sciences*, 10:1–6, 2015.

[16] Brian J Stucky, James P Balhoff, Narayani Barve, Vijay Barve, Laura Brenskelle, Matthew H Brush, Gregory A Dahlem, James DJ Gilbert, Akito Y Kawahara, Oliver Keller, et al. Developing a vocabulary and ontology for modeling insect natural history data: example data, use cases, and competency questions. *Biodiversity Data Journal*, 7, 2019.

[17] Archana Patel, Sarika Jain, Narayan C Debnath, and Vishal Lama. Inbiodiv-o: an ontology for indian biodiversity knowledge management. *arXiv preprint arXiv:2108.09372*, 2021.

[18] Gheorghe Tecuci, Dorin Marcu, Mihai Boicu, and David A. Schum. *Ontologies*, pages 155–173. Cambridge University Press, 2016.

[19] Natalya F. Noy and Deborah L. McGuinness. Ontology development 101: A guide to creating your first ontology. 01 2002.

[20] Jérôme Euzenat and Heiner Stuckenschmidt. Ontology matching, 06 2007.

[21] Fatima Ardjani, Djelloul Bouchiha, and Mimoun Malki. Ontology-alignment techniques: Survey and analysis. 7(11):67–78, 11 2015.

[22] Sameer M. Alrehaili, Mohammad Alqahtani, and Eric Atwell. A hybrid methods of aligning arabic qur’anic semantic resources. 03 2018.

[23] Xingsi Xue and Pei-Wei Tsai. Ecology and biodiversity ontology alignment for smart environment via adaptive compact evolutionary algorithm. *Frontiers in Plant Science*, 13:877120, 2022.

[24] Naouel Karam, Abderrahmane Khiat, Alsayed Algergawy, Melanie Sattler, Claus Weiland, and Marco Schmidt. Matching biodiversity and ecology ontologies: challenges and evaluation results. *The Knowledge Engineering Review*, 35, 2020.

[25] Alexandre Bento, Amal Zouaq, and Michel Gagnon. Ontology matching using convolutional neural networks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5648–5653, 2020.

[26] Sabrina Kirrane Sebastian Neumaier Axel Polleres Sabbir M Rashid Anisa Rula Lukas Schmelzeisen Steffen Staab Aidan Hogan, Claudia D'Amato. Knowledge graphs.

[27] Yucong Duan, Lixu Shao, and Gongzhu Hu. Specifying knowledge graph with data graph, information graph, knowledge graph, and wisdom graph. *IGI Global*, 6(2):10–25, 04 2018.

[28] Aidan Hogan, Claudia D'Amato, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Sabbir M Rashid, Anisa Rula, Lukas Schmelzeisen, and Steffen Staab. Knowledge graphs, 2020.

[29] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*, 2019.

[30] Aidan Hogan, Claudia D'Amato, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Sabbir M Rashid, Anisa Rula, Lukas Schmelzeisen, and Steffen Staab. Knowledge graphs, 2020.

[31] Deborah L. McGuinness. Owl web ontology language overview, 2004.

[32] Elin K Jacob. Ontologies and the semantic web. *Bulletin of the American Society for Information Science and Technology*, 29(4):19–19, 2003.

[33] Mike Dean, AT Schreiber, S Bechofer, FAH van Harmelen, Jim Hendler, Ian Horrocks, D MacGuinness, Peter Patel-Schneider, and Lynn Andrea Stein. Owl web ontology language reference. 2004.

[34] Grigoris Antoniou and Frank van Harmelen. Web ontology language: Owl. *Handbook on ontologies*, pages 91–110, 2009.

[35] Joel Sachs, Roderic Page, Steven J Baskauf, Jocelyn Pender, Beatriz Lujan-Toro, James Macklin, and Zacchaeus Comspon. Training and hackathon on building biodiversity knowledge graphs. *Research Ideas and Outcomes*, 5:e36152, 2019.

[36] Ramona L Walls, John Deck, Robert Guralnick, Steve Baskauf, Reed Beaman, Stanley Blum, Shawn Bowers, Pier Luigi Buttigieg, Neil Davies, Dag Endresen, et al. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLOS ONE*, 9(3):e89606, 2014.

[37] J. Neil Otte, John Beverley, and Alan Ruttenberg. Bfo: Basic formal ontology. *Applied Ontology*, 17:17–43, 2022.

[38] Archana Patel, Sarika Jain, Narayan C. Debnath, and Vishal Lama. Inbiodiv-o: An ontology for biodiversity knowledge management. *International Journal of Information Systems Modeling and Design*, 13:1–18, 2022.

[39] Cássia Trojahn, Renata Vieira, Daniela Schmidt, Adam Pease, and Giancarlo Guizzardi. Foundational ontologies meet ontology matching: A survey. *Semantic Web*, 13:685–704, 2021.

[40] Marcos Daniel Zárate, Germán Alejandro Braun, Pablo R. Fillottrani, Claudio Delrieux, and Mirtha Lewis. Bige-onto: An ontology-based system for managing biodiversity and biogeography data. *Applied Ontology*, 15:411–437, 2020.

[41] Inès Osman, Sadok Ben Yahia, and Gayo Diallo. Ontology integration: Approaches and challenging issues. *Elsevier BV*, 71:38–63, 07 2021.

[42] Donna Fritzsche, Michael Grüninger, Kenneth Bacławski, Mike Bennett, Gary Berg-Cross, Todd Schneider, Ram D. Sriram, Mark Underwood, and Andrea Westerinen. Ontology summit 2016 communique: Ontologies within semantic interoperability ecosystems. *IOS Press*, 12(2):91–111, 07 2017.

[43] Ramona L. Walls, John Deck, Robert Guralnick, Steve Baskauf, Reed Beaman, Stanley Blum, Shawn Bowers, Pier Luigi Buttigieg, Neil Davies, Dag Endresen, Maria Alejandra Gandolfo, Robert Hanner, Alyssa Janning, Leonard Krishtalka, Andréa Matsunaga, Peter Midford, Norman Morrison, Éamonn Ó Tuama, Mark Schildhauer, Barry Smith, Brian J. Stucky, Andrea Thomer, John Wieczorek, Jamie Whitacre, and John Wooley. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies, 03 2014.

[44] Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Oxford University Press*, 16(6):1069–1080, 04 2015.

[45] Fabio Ciotti and Francesca Tomasi. Formal ontologies, linked data, and tei semantics. *UMR ESPACE et UMR LISST*, (Issue 9), 09 2016.

[46] Hanna Köpcke and Erhard Rahm. Frameworks for entity matching: A comparison. *Elsevier BV*, 69(2):197–210, 02 2010.

[47] William W. Cohen, Pradeep Ravikumar, and Alan F. Karr. A comparison of string distance metrics for name-matching tasks. pages 73–78, 08 2003.

[48] Delip Rao, Paul McNamee, and Mark Dredze. Entity linking: Finding extracted entities in a knowledge base, 07 2012.

[49] Natalya F Noy. Semantic integration: a survey of ontology-based approaches.

[50] Flor K. Amanqui, Kleberson J. Serique, Silvio Domingos Cardoso, Jose L. Dos Santos, Andréa Corrêa Flôres Albuquerque, and Dilvan de Abreu Moreira. Improving biodiversity data retrieval through semantic search and ontologies. 08 2014.

[51] Thomas Berg and Peter N. Belhumeur. How do you tell a blackbird from a crow? 12 2013.

[52] Ernesto Jiménez-Ruiz, Asan Agibetov, Jiaoyan Chen, Matthias Samwald, and Valerie Cross. Dividing the ontology alignment task with semantic embeddings and logic-based modules, 01 2020.

[53] Robert Hoehndorf, Michel Dumontier, Anika Oellrich, Dietrich Rebholz-Schuhmann, Paul N. Schofield, and Georgios V. Gkoutos. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *Public Library of Science*, 6(7):e22006–e22006, 07 2011.

[54] Ramona Walls, Robert Guralnick, John Deck, Adam Buntzman, Pier Luigi Buttigieg, Neil Davies, Michael Denslow, Rachel E. Gallery, Jacob Parnell, David Osumi-Sutherland, Robert Robbins, Philippe Rocca-Serra, John Wieczorek, and Jie Zheng. Meeting report: advancing practical applications of biodiversity ontologies. *Springer Science+Business Media*, 9(1), 12 2014.

[55] Barry Smith. Classifying processes: An essay in applied ontology. *Wiley*, 25(4):463–488, 11 2012.

[56] Luiz Gadelha, Pedro C. de Siracusa, Eduardo Dalcin, Luís Alexandre Estevão da Silva, Douglas A. Augusto, Eduardo Krempser, Helen Michelle de Jesus Affe, Raquel L. Costa, Maria Luiza Mondelli, Pedro Milet Meirelles, Fabiano L. Thompson, Márcia Chame, Artur Ziviani, and Marinez Ferreira de Siqueira. A survey of biodiversity informatics: Concepts, practices, and challenges. *Wiley*, 11(1), 11 2020.

[57] John Deck, Robert Guralnick, Ramona Walls, Stanley Blum, Melissa Haendel, Andréa Matsunaga, and John Wieczorek. Meeting report: Identifying practical applications of ontologies for biodiversity informatics. *Springer Science+Business Media*, 10(1), 05 2015.

[58] Janisha Patel, Dennis A. Dean, Charles H. King, Nan Xiao, Soner Koc, Ekaterina Minina, Anton Golikov, Phillip Brooks, Robel Kahsay, Rahi Navelkar, Manisha Ray, Dave Roberson, Chris Armstrong, Raja Mazumder, and Jonathon Keeney. Bioinformatics tools developed to support biocompute objects. *University of Oxford*, 2021, 01 2021.

[59] Florent Mazel, Matthew W. Pennell, Marc W. Cadotte, Sandra Dıaz, Giulio Valentino Dalla Riva, Richard Grenyer, Fabien Leprieur, Arne Ø. Mooers, David Mouillot, Caroline M. Tucker, and William D. Pearse. Prioritizing phylogenetic diversity captures functional diversity unreliably. *Nature Portfolio*, 9(1), 07 2018.

[60] Shannon D. Bower, Jacob W. Brownscombe, Kim Birnie-Gauvin, Chris K. Elvidge, Andrew D. Moraga, Ryan Pusiak, Eric D. Turenne, Aaron J. Zolderdo, Steven J. Cooke, and Joseph Bennett. Making tough choices: Picking the appropriate conservation decision-making tool. *Wiley*, 11(2), 11 2017.

[61] Houda El Bouhissi, Sarika Jain, Narayan C. Debnath, and Vishal Lama. Inbiodiv-o: An ontology for indian biodiversity knowledge management, 01 2021.

[62] Joshua S. Madin, Shawn Bowers, Mark Schildhauer, and Matthew B. Jones. Advancing ecological research with ontologies, 03 2008.

[63] Anne Frances, Leah M. Oliver, and Kathy Goodin. Conservation implications of taxonomic intelligence: A case study of trillium. *Pensoft Publishers*, 4, 10 2020.

[64] Malcolm Barcant. *Butterflies of Trinidad and Tobago*. 1971.

[65] Katherine LeVan. Specimens in a broader context: The national ecological observatory network and the extended specimen. *Biodiversity Information Science and Standards*, 4:e59208, 2020.

[66] Protected planet: Trinidad and tobago (tto), 2023. Accessed: 2023-09-17.