Semantic Data Integration for Forest Observatory Applications

Naeima Hamed

School of Computer Science and Informatics
Cardiff University

This dissertation is submitted for the degree of
Doctor of Philosophy

June 2024

Abstract

The populations of endangered species, such as African and Asian elephants, are declining due to habitat loss, fragmentation, and poaching. Driven by the ivory trade, poaching involves the unlawful killing of animals, posing a significant threat to elephant populations. Due to funding shortages in wildlife conservation, analysing research data has emerged as a cost-effective solution for decision-making in protecting wildlife species. Wildlife research data are typically collected on a project-specific basis, leading to the creation of data silos. Addressing key conservation questions often requires unified access to diverse data sources. For example, accessing elephants' tracked movements alongside environmental data during the dry or wet season can help predict whether they are heading towards locations that expose them to poachers.

This research introduces a novel approach that employs semantic web technologies to integrate heterogeneous wildlife data from the forests of Sabah in Malaysian Borneo. A review of Open Data Observatories and their data management methods identified the Semantic Web as an effective approach to breaking wildlife research data silos. Consequently, the Forest Observatory Ontology (FOO) was developed to standardise sensor-monitored wildlife data for integration. FOO was populated with four heterogeneous wildlife datasets to construct knowledge graphs. Predictive models derived from these knowledge graphs were used to predict elephants' geo-locations and poaching likelihood, providing a proactive tool for conservationists. To extend the research, the generalisation of the methodology to different domains was explored by developing and populating another ontology for Internet of Things (IoT) data marketplaces, enabling on-demand data purchasing.

This doctoral research contributes to wildlife data management by analysing Open Data Observatories to identify optimal approaches for integrating data. It develops the Forest Observatory Ontology (FOO) and its associated knowledge graphs to standardise and unify wildlife data generated by sensors. Using the constructed knowledge graphs, the research creates predictive models for poaching through deep learning and semantic reasoning.

Contents

Abstract **i**

List of Tables **ix**

List of Figures **xi**

Acknowledgement **xiii**

List of Publications **xv**

1 Introduction **1**
1.1 Motivation 1
1.2 Problem Statement 3
1.3 Research Questions 4
1.4 Research Contributions 8
1.5 Thesis Structure 8

2 Literature Review **11**
2.1 Introduction 11
2.2 Open Data 13
2.2.1 Open Data Principles 13
2.2.2 Open Data Sources 15
2.3 Research Method 15
2.3.1 Search Plan 16
2.3.2 Observatories Selection Process 16
2.4 Open Data Observatories 17
2.4.1 Terrestrial Ecosystem Research Network (TERN) 19
2.4.2 Channel Coastal Observatory (CCO) 19
2.4.3 Urban Observatory Project (UOP) 20

2.4.4 Global Forest Watch (GFW) . . . 21
2.4.5 Global Earth Observation System of Systems (GEOSS) . . . 21
2.4.6 Earth Observing System Data and Information System (EOSDIS) . . . 22
2.4.7 Grow Observatory (GROW) . . . 22
2.4.8 Tsunami Observatory . . . 23
2.4.9 Southampton Data Observatory (SDO) . . . 24
2.4.10 National Ecological Observatory Network (NEON) . . . 24
2.4.11 India Urban Observatory (IUO) . . . 25
2.4.12 The Finnish Ecosystem Observatory (FEO) . . . 25
2.4.13 The Open Forest Observatory (OFO) . . . 26
2.5 Data Themes, Sources, Processing and Visualisation . . . 26
2.5.1 Urban Data Theme . . . 27
2.5.2 Non-urban Data Theme . . . 28
2.5.3 Data Themes Comparison . . . 29
2.5.4 Data Sources . . . 30
2.5.5 Data Processing . . . 31
2.5.6 Data Visualisation . . . 33
2.6 Research Challenges . . . 33
2.6.1 Data Integration . . . 34
2.6.2 Data Quality . . . 34
2.6.3 Data Provenance . . . 35
2.6.4 Data Privacy . . . 36
2.6.5 Takeaways . . . 38
2.7 Data Management Approach for Wildlife Data . . . 39
2.7.1 Data Integration for Wildlife Applications . . . 39
2.7.2 Semantic Modelling for Wildlife Data . . . 41
2.7.3 Wildlife Ontologies . . . 42
2.7.4 Ontology Development Methodologies . . . 42
2.8 Knowledge Graphs . . . 43
2.8.1 Knowledge Graphs Creation Methodologies . . . 44
2.8.2 Ontologies for Knowledge Graphs . . . 45
2.8.3 Knowledge Graphs for Data Modelling . . . 45
2.8.4 Knowledge Graphs for Crime Prediction . . . 47
2.8.5 Wildlife Crime prediction . . . 48
2.9 Summary . . . 52

3 Forest Observatory Ontology Data Store (FooDS) 55
3.1 Introduction . . . 55
3.2 Forest Observatory Ontology (FOO) . . . 57
3.2.1 Ontology Requirements . . . 58
3.2.2 Ontology Implementation . . . 64
3.2.3 Ontology Evaluation . . . 72
3.2.4 Ontology Publication and Maintenance . . . 77
3.3 Forest Ontology Observatory Data Store (FooDS) . . . 78
3.3.1 Soil RDF Graph . . . 79
3.3.2 Vegetation RDF Graph . . . 81
3.3.3 GPS Collar RDF Graph . . . 82
3.3.4 Camera Trap Images RDF Graph . . . 82
3.3.5 Semantic Data Integration . . . 83
3.4 FooDS Evaluation . . . 87
3.4.1 Domain Experts Evaluation . . . 88
3.4.2 Use Case 1: Elephants spending time together . . . 88
3.4.3 Use Case 2: Salt licks locations . . . 89
3.4.4 Use Case 3: Rescuing the injured elephant . . . 89
3.4.5 Results . . . 91
3.4.6 Discussion . . . 92
3.5 Summary . . . 94

4 Leveraging FooDS for Predicting Wildlife Poaching 95
4.1 Introduction . . . 95
4.2 Elephant GPS Collar Knowledge Graph RDF . . . 96
4.3 Geo-location Prediction with Deep Learning . . . 97
4.3.1 Data Extraction and Preprocessing . . . 98
4.3.2 Architecture and Training . . . 98
4.4 Rule-based Reasoning for Poaching Prediction . . . 99
4.5 Results . . . 103
4.5.1 Geo-locations Prediction Result . . . 104
4.5.2 Evaluation . . . 104
4.5.3 Data Preprocessing . . . 104
4.5.4 Models and Results . . . 105
4.6 Discussion . . . 106
4.7 Summary . . . 108

5 Extending FooDS's Semantic Web Framework to Data Marketplaces 109
5.1 Introduction . . . 109
5.2 Motivation and the Problem Definition . . . 111
5.2.1 Design Principles and Architecture . . . 113
5.3 Data Marketplace Design . . . 114
5.3.1 Data Model . . . 116
5.3.2 Ontology Requirements . . . 116
5.3.3 Ontology Analysis . . . 117
5.3.4 Ontology Implementation . . . 119
5.3.5 Ontology Evaluation . . . 120
5.3.6 Experimentation Plan . . . 120
5.4 Evaluation . . . 122
5.4.1 Rule-based Reasoning . . . 125
5.4.2 Evaluation One . . . 126
5.4.3 Evaluation Two . . . 128
5.4.4 Evaluation Three . . . 130
5.4.5 Results . . . 130
5.4.6 Cost of Adapting Linked Data . . . 133
5.4.7 Use Cases . . . 134
5.5 Discussion . . . 137
5.6 Summary . . . 139

6 Conclusion 141
6.1 Research Questions and Contributions . . . 142
6.2 Research Novelty . . . 143
6.3 Future Work . . . 144
6.4 Concluding Remarks . . . 146

Bibliography 147
.1 Appendix I: Methodology Details . . . 169
.2 Ontology Requirements Specification Document (ORSD) . . . 171
.2.1 Purpose . . . 171
.2.2 Scope . . . 171
.2.3 Implementation Language . . . 171
.2.4 Intended End-Users . . . 171
.2.5 Intended Uses . . . 171
.2.6 Ontology Requirements . . . 171

.3 Competency Questions and their formulated SPARQL Queries 172
.4 Appendix II: Forest Observatory Ontology (FOO) 238

List of Tables

2.1 A Comparison of Open Data Principles 13
2.2 Open Data Observatories Data types 27
2.3 Open Data Observatories Data Themes 29
2.4 Common Themes and Differences 30
2.5 Newcastle Urban Observatory Parameters 31
2.6 Open Data Observatories Data Source 32
2.7 Open Data Observatories Strengths and limitations 40
2.8 Ontology development Methodologies 43
2.9 Benefits of Ontology-based Knowledge Graphs 45

3.1 Overview of Participant Feedback Themes 64
3.2 Competency Questions (CQs) 65
3.3 Competency Questions (CQs)-Continued 66
3.4 Natural Language Statements (NLSs) 67
3.5 FOO's main classes 73
3.6 SPARQL Queries Performance 77
3.7 Soil data descriptive analysis 81
3.8 Lianas data descriptive analysis 82
3.9 Usability Study Results 87
3.10 Competency Questions (CQs) Response Time 91

4.1 Summary Statistics of GPS Data and Related Metrics 97
4.2 Model Evaluation 106

5.1 Possible Scenarios 116
5.2 Proposed data model 119
5.3 Data Model RDF Graph Size (KB) 125

1 Natural Language Statements 232

List of Figures

1.1 Wildlife Data generated by a Sensor 6
1.2 Thesis Structure 10

2.1 Inclusion and Exclusion Criteria 17
2.2 Open Data Observatories Timeline 18
2.3 Transport Data by Open Data Observatories 28
2.4 Newcastle Urban Observatory Parameters Count 31
2.5 Research Challenges 37
2.6 Wildlife Crime Mitigation 51

3.1 Ontology Development Phases 58
3.2 Participants' Demographic Information 60
3.3 Ontology Development Activities 61
3.4 Ontology Conceptual Diagram 74
3.5 Ontology Conceptual Diagram 75
3.6 FOOPS! Results 76
3.7 Ontology and Knowledge Graphs Shared Concepts 79
3.8 Knowledge Graph Construction UML 79
3.9 Soil Knowledge Graph 80
3.10 Vegetation RDF Graph 81
3.11 GPS Collar RDF graph 83
3.12 Camera Trap Images RDF Graph 83
3.13 Semantic Data Integration 85
3.14 System Design 86
3.15 Usability Ratings Box Plot 88
3.16 Competency Questions (CQs) Response Time 92

4.1 Elephant Seri GPS tracker Knowledge Graph 97
4.2 PoachNet Architecture 99

4.3 Intersection between an oil palm plantation and an elephant's movements . 102
4.4 Oil palm plantation and elephant's movements 102
4.5 Models Evaluation 106

5.1 Urban Data Marketplace 111
5.2 Participants Background 114
5.3 Participants Demographics 114
5.4 Semantically Enhanced IoT Data Marketplace Architecture 115
5.5 Neon Methodology 117
5.6 Urban Data Exchange Ontology Classes 119
5.7 Urban Data Exchange Ontology Instances 121
5.8 Experimental Setup 122
5.9 Comparison between Data Model Sizes(KB) 125
5.10 Evaluation 1 visualisation 127
5.11 Evaluation 1 boxplot 127
5.12 Evaluation 1 No-Rule distribution plot 127
5.13 Evaluation 1 SWRL distribution plot 127
5.14 Evaluation 2 visualisation 129
5.15 Evaluation 2 boxplot 129
5.16 Evaluation 2 No-Rule distribution plot 129
5.17 Evaluation 2 SWRL distribution plot 129
5.18 Evaluation 3 Setup 131
5.19 Evaluation 3 visualisation 131
5.20 Evaluation 3 boxplot 131
5.21 Evaluation 3 No-Rule distribution plot 131
5.22 Evaluation 3 SWRL distribution plot 131
5.23 Evaluation Comparison 132
5.24 Descriptive Analysis 133
5.25 Evaluation Statistical Analysis 133
5.26 Evaluations query average time comparison 133

1 FOOPS! Score 266

Acknowledgement

I sincerely thank my lead supervisor, Dr. Charith Perera, for his guidance, patience, and encouragement, which have been crucial to my success. A massive thanks to Professor Omer Rana for his wisdom and support, which have pushed me to improve and achieve more—it has been an honour to work with you.

I'm truly thankful to Dr. Pablo Orozco Ter Wengel for his mentorship and consistent support throughout this journey. I extend my deepest gratitude to Professor Benoit Goossens for his thoughtful feedback, enduring support, and vital contributions.

I would like to thank the staff and researchers at the Danau Girang Field Centre (DGFC) for their support, hospitality, and commitment during my fieldwork. Your dedication to conservation has been truly inspiring.

I extend my heartfelt thanks to my examiners and chair, Professor John Domingue, Dr. Alia Abdelmoty, and Dr. Georgios Theodorakopoulos, for providing an incredible and unforgettable experience. *I am deeply grateful for the time and effort you dedicated to examining my thesis and for your thoughtful insights and feedback.*

To my friends, colleagues, and fellow student representatives in Computer Science and the IoT Garage, thank you for your hard work and support—it has been a pleasure working with such talented individuals.

To my parents, your love and support have been the foundation of everything I have achieved. I hope I have made you proud. To my husband, your patience and encouragement have been my strength, and to my wonderful children, your smiles and laughter have kept me going even on the hardest days.

Lastly, I am grateful to the academics and staff at Cardiff University's Schools of Computer Science & Informatics and Mathematics. The support and resources you have provided have been essential to my growth and success.

List of Publications

This thesis contains work presented in the following publications:

Journal paper Naeima Hamed, Andrea Gaglione, Alex Gluhak, Omer Rana, and Charith Perera. 2023. Query Interface for Smart City Internet of Things Data Marketplaces: A Case Study. ACM Trans. Internet Things 4, 3, Article 19 (August 2023), 39 pages. https://doi.org/10.1145/3609336.

Book chapter Naeima Hamed, Omer Rana, Pablo Orozco-terWengel, Benoît Goossens, and Charith Perera. (2023). FOO: An Upper-Level Ontology for the Forest Observatory. In: Pesquita, C., et al. The Semantic Web: ESWC 2023 Satellite Events. ESWC 2023. Lecture Notes in Computer Science, vol 13998. Springer, Cham. https://doi.org/10.1007/978-3-031-43458-7_29.

Journal paper Naeima Hamed, Omer Rana, Pablo Orozco-terWengel, Benoît Goossens, and Charith Perera. 2024. A Comparison of Open Data Observatories. J. Data and Information Quality Just Accepted (November 2024). https://doi.org/10.1145/3705863.

Journal paper Naeima Hamed, Omer Rana, Pablo Orozco-terWengel, Benoît Goossens, and Charith Perera. 2024. FooDS: Ontology-based Knowledge Graphs for Forest Observatories. ACM Journal on Computing and Sustainable Societies Just Accepted (November 2024). https://doi.org/10.1145/3707637.

Journal paper Naeima Hamed, Omer Rana, Pablo Orozco-terWengel, Benoît Goossens, and Charith Perera. 2024. PoachNet: Predicting Poaching Activities through Integrating Heterogeneous Opportunistic Wildlife Data (Under Journal Review).

Chapter 1

Introduction

1.1 Motivation

Many populations of endangered species, such as elephants, are declining due to habitat loss, fragmentation, and poaching [157, 106, 3]. Also, conflicts between humans and elephants sometimes escalate, leading to injuries and death on both sides. Forest fragmentation occurs mainly due to human activities such as logging, agricultural expansion, infrastructure development, and urbanisation. These activities break large, contiguous forests into smaller, isolated patches, leading to habitat loss and limiting elephants to smaller and more vulnerable populations. Poaching, on the other hand, refers to the unlawful killing and capturing of animals. Despite continuous conservation efforts to curb and combat poaching, this crime keeps happening [153, 70, 252].

Both African and Asian elephant species are under constant threat from poaching due to the profitable ivory trade. IUCN (International Union for Conservation of Nature) recognises two distinct species of African elephants, the savannah elephant (Loxodonta africana), classified as endangered, and the forest elephant (Loxodonta cyclotis), listed as "critically endangered" on the IUCN Red List [93, 33]. Since the early 2000s, the African elephant population has experienced significant declines, with approximately 415,000 individuals living in 37 African countries [5]. The Asian elephant (Elephas maximus) species is distributed across 13 countries in Asia and is also classified as endangered, with a decreasing population estimated to be between 36,000 and 50,000 individuals [205, 157].

Elephants are the ecosystem engineers who reshape forests with their travelling and feeding habits [184]. As they move between the forest corridors, tall trees are knocked down, creating grasslands and savannas that allow other species to thrive. As mega-herbivores, the germination process taking place in their digestive systems releases dung that is rich in seeds, enabling widespread seed dispersal [142, 126, 25]. From an economic point of view,

elephants are one of the pillars of wildlife tourism that yields decent income and opens jobs for many people in some developing countries [182]. Culturally, elephants are revered in many societies. They are featured prominently in religion, mythology, and national symbols [109].

Wildlife conservation bodies in developing countries often face resource limitations, understaffing, and chronic underfunding [193]. Early research in the 2000s by James et al. [131] pointed out significant imbalances in conservation funding, with the most biodiverse regions receiving minimal financial support. In the mid-2000s, Balmford et al. [19] identified bureaucratic inefficiencies that further limited these scarce resources. Subsequent studies by McCarthy et al. [171] showed that funding needs to increase by as much as five to ten times the current levels to manage protected areas effectively. Waldron et al. [257] later confirmed these financial constraints in developing countries, stressing the mismatch between the amount of money invested and the ecological value of these investments. If wildlife conservation bodies were funded similarly to police forces, they could recruit more personnel (wildlife practitioners and rangers) to expand the area coverage of ranger patrols and enhance on-the-ground enforcement. Sufficient funds can equip rangers with necessary health and safety supplies, guaranteed access to reasonable shelter, fresh food, clean drinking water, and reliable transportation [229].

The shortage of funding for wildlife conservation has led to a shift towards analysing wildlife research data as a more cost-effective substitute for traditional, resource-heavy methods. Urbano et al. [247] highlighted the importance of a scientifically informed approach to decision-making, particularly when conservation resources are limited. Wildlife research data are typically gathered through various methods, including field surveys, Global Positioning System (GPS) tracking, motion-activated trail cameras, and low-cost sensors. Field surveys involve observing animal sightings, tracking signs, and counting animals at crucial sites. Aerial surveys, conducted using drones allow researchers to spot and count wildlife from above. GPS collars around elephants' necks enable park officials to track their individual or herd locations [135, 128, 62].

Motion-activated trail cameras, triggered by movement, autonomously capture images or videos of wildlife in their natural environments without human interference. Likewise, airborne sensors such as digital cameras, Light Detection and Ranging (LIDAR), and imaging spectrometers can be attached to aircraft to carry out aerial surveys and map wildlife habitats and populations. Acoustic sensors [220, 117, 18] can detect the sound of gunshots and alert rangers to the location of potential poaching incidents. A study by Zwerts et al. [283] found that data-driven approaches using passive acoustic monitoring and motion-activated trail cameras were more cost-effective and efficient than traditional human observer surveys,

especially in remote and hard-to-access areas. Developing smaller, lower-cost GPS loggers and other sensors allows conservation efforts to scale monitoring and data collection across larger areas, even with limited budgets. This scalability is challenging to achieve with boots-on-the-ground ranger patrols alone.

1.2 Problem Statement

Wildlife research data are typically collected on a project-by-project basis and often exist in silos. Data silos happen due to independent data management, analysis, and storage by different research activities. This data isolation hinders collaboration as groups work independently, limiting the ability to answer key questions beyond a specific project [167].

Wildlife research data management has recently revolved around systems that focus on gathering research data from disparate sources [243, 183, 266]. These systems include Open Data portals like *data.gov.uk*, which centralise government and institutional data for public access [161]. Open Data Observatories such as *tern.org.au* monitor and analyse domain-specific datasets for trends, and generalist repositories such as *zenodo.org* archive diverse scholarly outputs, supporting interdisciplinary research and increasing the visibility of academic work.

However, data integration methods in some of these systems often lack automated logical connections between data entities and their relationships. Although various data can be freely offered by such environments, *to the best of my knowledge, no system so far has offered queryable, integrated, and linked wildlife data for the area and data types used in this research.* Various data-driven methods were employed in wildlife conservation [97, 98, 41]. These methods used Geographic Information Systems (GIS), machine learning, artificial intelligence, and predictive analytics to understand species migration patterns, formulate protective measures for at-risk species, optimise conservation resources and detect poaching intentions. Yuan et al. [278] proposed a study optimising land use and patrol routes using spatial data and mathematical optimisation to allocate conservation resources strategically. Complementing this, Park et al. [199] introduced a behavioural model-based anti-poaching engine that simulates poacher actions to refine patrol strategies dynamically, integrating predictive behaviour modelling to adapt to evolving poaching patterns. Fang et al. [83] used historical poaching data and game theory to predict poaching hotspots and incorporated environmental features to improve patrol efficiency. Predictive and spatial strategies were deemed helpful in deploying resources where they are most needed.

Furthermore, a study by Gurumurthy et al. [113] combined sparse data with domain expert knowledge (rangers) to tackle poaching. Zafra-Calvo et al. [279] established a connection

between the locations of elephant remains and their closeness to roads and protected regions. Their research findings favoured the inclusion of diverse data into conservation approaches. In parallel, Chen et al. [43] adopted a multimodal data integration approach for patrol planning that synthesises geographical, temporal, and specific incident data, thereby broadening the scope of data utility. Moreover, such data usage is theoretically promoted by Neil et al. [184], whose agent-based modelling simulates different strategies, assessing their performance in controlled environments. Further enhancing the data foundation of these models, Rivera [221] addresses the challenge of unreliable datasets by employing longitudinal analysis techniques to clean data inputs. Cleaning these unreliable datasets improved their quality and made them fit for random forest algorithms employed by Jin et al. [133] to predict poaching. The accuracy of these predictions directly benefits from the enhanced data quality provided by Rivera's techniques. Kar et al. [138] focused on modelling adversary behaviour to predict poaching risks. Their work not only predicts where poaching might occur but also suggests proactive strategies to mitigate these risks, seamlessly tying into the framework established by the predictive analytics of Jin et al.

At the corporate level, the World Wildlife Fund's Elephant Conservation Unit in Malaysia uses Global Positioning System (GPS) collars, professionally fitted on several elephants in the forests of Sabah, primarily to reduce human-elephant conflicts. This type of data and other wildlife information hold untapped potential for further in-depth analysis and knowledge discovery. Chibeya et al. [48] as such demonstrated that collaring data, combined with environmental factors, can predict elephant locations during the wet season. Although these technologies have enhanced animal movement tracking, the underlying reasons for these movements remain underexplored. There is also a gap in linking diverse wildlife data with context and advanced algorithms to achieve more accurate poaching prediction.

1.3 Research Questions

This doctoral thesis proposes a novel approach for integrating diverse wildlife data and uses these integrated data to predict poaching. This research collaborated with the Danau Girang Field Centre (DGFC) (danaugirang.com), a research and education facility located in the heart of Sabah, Malaysia, within the Lower Kinabatangan Wildlife Sanctuary. This sanctuary, spanning approximately 270 km² and situated between E 118°00' - 118°50', N 5°20' - 5°50', features a tropical rainforest climate and is home to a variety of endangered species (e.g., Bornean elephants (Elephas maximus), orangutans (Pongo pygmaeus), and Sunda pangolins (Manis javanica)). DGFC focuses on conserving biodiversity and ecosystems in the region through scientific research. The centre studies how wildlife adapts to fragmented landscapes

caused by deforestation and human activity. In addition to its research activities, DGFC provides educational programmes, including internships and field courses for university students, aiming to train the next generation of conservation scientists and increase awareness of environmental issues. The main focus of research is on the situation of Bornean elephants in the Lower Kinabatangan region of Sabah, Malaysian Borneo, which remains a critical concern as they face the persistent threat of poaching, human-elephant conflict and habitat loss. Illegal killings and injuries to elephants sometimes occur due to conflicts with farmers. Elephants invade human settlements such as oil palm plantations, causing serious harm to humans, properties, and machinery. Moreover, elephants often fall victim to snare traps set for wild boar and deer in forest areas *near oil palm plantations*, like the Kinabatangan floodplain. Since 2010, it's estimated that 20% of Bornean elephants have been injured by these snares [7].

The overarching research question is : *Can a 'Linked Data Store' be developed to answer questions supporting wildlife research and conservation activities in the wild?*

To reach an answer to the overarching research question, three sub-questions emerged:

Research Question 1 (RQ1): *Can an effective data management approach be developed to integrate heterogeneous wildlife data from disparate sources?*

To find an effective data management approach, a literature review was conducted. Thirteen open Data Observatories were selected, examined, and compared. Open Data Observatories are online data platforms that integrate heterogeneous data from disparate sources. The comparison was based on their data types, domain coverage, accessibility, and usability. Their data management approaches were compared and analysed to identify and adopt a suitable approach for this research. Furthermore, significant technical and intellectual challenges were identified. The findings from the literature review guided the recommendation to employ semantic web technologies as an effective data management approach for this research. Semantic web technologies have the capability to integrate heterogeneous data from disparate sources.

To illustrate the concept, semantic web technologies use ontologies to model the scenario of an elephant fitted with a GPS tracking collar as a Resource Description Framework (RDF) graph consisting of relationships such as (*subject, predicate, object)*, for example, as shown in Figure 1.1, this graph contains entities and relationships like the gPS_tracking_Collar is an instance of the class *Sensor* (GPS_tracking_Collar $\in$ Sensor). The *Sensor* made an observation (*Sensor* $\xrightarrow{\text{made}}$ Observation). The gPS_tracking_Observation is an instance of the

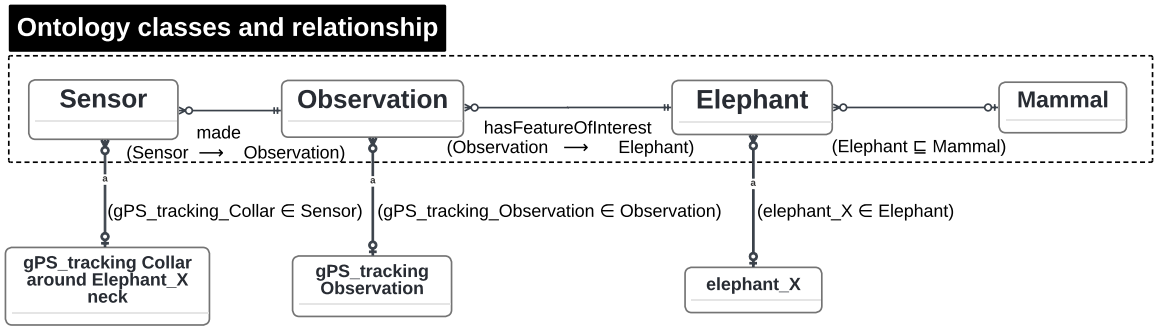


Figure 1.1 Wildlife data generated by an Internet of Things (IoT) sensor, modelled as an RDF graph using ontologies.

class *Observation* (GPS_tracking_Observation $\in$ Observation).
Class Observation has the feature of interest, which is class Elephant (Observation $\overset{\text{hasFeatureOfInterest}}{\longrightarrow}$ Elephant). The so called elephant_X is an instance of the class *Elephant* (elephant_X $\in$ Elephant), and class *Elephant* is a subclass of *Mammal* (Elephant $\sqsubseteq$ Mammal).

Ontologies [111] are structured frameworks that describe the types, properties, and interrelationships of concepts within a specific domain. They can define and infer the logical connections between concepts (data entities) and their relationships.

Accordingly, a further review of the literature related to wildlife data management was conducted, focusing on the application of semantic web technologies in modelling wildlife data and comparing different methodologies. The development and advantages of using ontologies and knowledge graphs were briefly explored, respectively. Existing studies on knowledge graphs in predictive modelling and crime prediction were also examined alongside the advancements in wildlife crime prediction techniques.

Knowledge graphs [118] represent a way of structuring and integrating knowledge based on relationships between entities (such as objects, individuals, concepts, or events), enabling humans and machines to interpret and process the interconnected information. Ontology-based knowledge graphs focus on developing semantic relationships in data. These relationships form meaningful connections between concepts in a particular domain, enabling an understanding and interpretation of how these concepts relate to each other. Semantic web technologies enable precise querying, complex relationship analysis, semantic consistency, and interoperability between different systems and data formats. Moreover, the reasoning capabilities can infer implicit knowledge that is not overtly specified within the data.

Research Question 2 (RQ2): *Can a 'Linked Data Store' be developed to answer questions supporting wildlife research and conservation activities?*

To address RQ2, an ontology named the Forest Observatory Ontology (FOO)[1] and its online documentation[2] were developed to standardise wildlife data generated by sensors. Then, a resource website was built for FOO and its knowledge graphs, and created an analytical dashboard, executable notebook and embedded a conservation AI Chatbot (as a proof of concept) to remotely query, visualise, and analyse four distributed wildlife knowledge graphs based on FOO.

The **Forest Observatory Ontology (FOO)** is a novel ontology built from data collected from wildlife research. It reuses entities from established ontologies to unify the Internet of Things (IoT) and wildlife concepts (biodiversity, conservation biology, habitat fragmentation, and endangered species management). To break wildlife data silos, FOO was populated or instantiated with four heterogeneous datasets transformed into Resource Description Framework (RDF) to produce ontology-based knowledge graphs, named the**Forest Observatory Ontology Data Store (FooDS)**[3]. To access and use FooDS, an interface was created to enable authorised users to script granular *(*SPARQL) search queries and retrieve instant answers to questions from integrated and remotely located datasets.

RQ2 contributions (FOO and FooDS) provide a novel (modular) approach to manage and integrate wildlife data to answer questions that support wildlife research and conservation activities.

Research Question 3 (RQ3): *Can prediction models be developed to predict poaching crimes by using the developed 'Linked Data Store'?*

To build the predictive models, data extracted from FooDS (i.e., ontology-based knowledge graph) were used to train a deep learning model to predict *tracked Bornean elephants* geo-locations. Then, semantic reasoning incorporated into FooDS was used to predict poaching likelihood based on contextual data. This chapter aims to augment bioscientists and conservationists in improving poaching prediction using modular and scalable predictive data model enriched with semantic reasoning.

Research Question 4 (RQ4): *Can the Linked Data Store's semantic web data management approach be generalised to another domain for various purposes?*

To address RQ4, FooDS's semantic web data management approach was generalised for use in the IoT data marketplace. In traditional data marketplaces, data are often sold as entire

[1] https://w3id.org/def/foo#
[2] https://w3id.org/def/fooDocs
[3] https://w3id.org/def/fooDS

datasets. This approach can be expensive and inefficient, as consumers may not require the entirety of the dataset but only specific observations or subsets of the data. To address this issue, FooDs's generalised approach enables consumers, particularly SMEs, to customise their data purchasing requests.

1.4 Research Contributions

The contributions of this thesis are as follows:

1. Contribution 1 (C1): A literature review was conducted, selecting and comparing thirteen Open Data Observatories. This was followed by an examination of semantic web technologies for wildlife data, evaluating methodologies and exploring the benefits of ontologies and knowledge graphs in predictive modelling and wildlife crime detection.
2. Contribution 2 (C2): Built the Forest Observatory Ontology (FOO) and populated it with four RDF graphs, developing the Forest Observatory Ontology Data Store (FooDs).
3. Contribution 3 (C3): Employed FooDs to build predictive models that forecast future elephant geo-locations and infer poaching.
4. Contribution 4 (C4): Generalised FooDs's semantic web data management approach to the IoT data marketplace. Built ontology-based knowledge graphs facilitating unique on-demand data offers. This contribution enabled IoT data buyers to customise their data purchasing requests.

1.5 Thesis Structure

What remains from this thesis is structured as follows:

Chapter 2 (Literature Review): This chapter compares thirteen Open Data Observatories and their data management approaches. It investigated their aims, design, types of data, and research challenges that influence the implementation of these observatories, outlining some advantages and limitations for each one and recommending areas for improvement. One of the findings of this review recommends semantic web technologies for effective data management. Subsequently, further review was conducted, focusing on the application of semantic web technologies to wildlife data, evaluating different methodologies, and exploring the benefits of using ontologies and knowledge graphs in predictive modelling and wildlife crime prediction. This chapter addresses RQ1, achieves C1.

Chapter 3 (Forest Observatory Ontology Data Store (FooDS)): introduces the Forest Observatory Ontology (FOO) and its development lifecycle and proposes FooDS (i.e., ontology-based knowledge graphs) the Linked Data Store. Forest Observatory refers to online platforms that aggregate, curate, integrate, store, and analyse heterogeneous wildlife data for effective forest monitoring. However, integrating such data from disparate sources can be challenging due to independent data management systems. This chapter proposes a novel approach for integrating diverse wildlife data into Forest Observatories. It employs knowledge graphs built on ontologies, enabling instant question-answering and inferences across isolated data sources. The Forest Observatory Ontology (FOO) standardised entities in the IoT and wildlife research data. Then, FOO was populated with four semantically modelled wildlife datasets. The result is the Forest Observatory Ontology Data Store (FooDS), containing over six million triples of heterogeneous wildlife data. Open-source tools, domain experts' validation, and use cases were used to evaluate the structure of FOO and the usability of FooDS. This Linked Data Store answers questions to support wildlife research and conservation activities. This chapter addresses RQ2 and achieves C2.

Chapter 4 (Leveraging FooDS for Predicting Wildlife Poaching): This chapter proposes PoachNet, a predictive tool that forecast elephants' geo-locations and poaching likelihood. PoachNet extracts granular data from FooDS, applies deep learning models, evaluates them, and returns accurate predictions to its database (triple-store). Output datasets can include diverse entities such as elephant GPS observations, soil conditions, and vegetation types. Semantic reasoning is then applied to the dataset to infer poaching. PoachNet equips conservationists with a useful tool to predict future elephant locations and poaching likelihood. This chapter addresses RQ3 and achieves C3.

Chapter 5 (Extending FooDS's Semantic Web Framework to Data Marketplaces): generalises FooDs's semantic web data management approach to a different domain. An ontology was developed with expert input and reuse, populated with datasets from various sensors to construct knowledge graphs and apply reasoning. This allowed them to acquire granular data records tailored to their needs from various data sources or providers, instead of being forced to purchase entire datasets, much of which may be irrelevant or unnecessary for their purposes. This chapter addresses RQ4 and achieves C4.

Chapter 6 (Conclusion) concludes the thesis by reminding the reader with the research questions and their corresponding contributions, research novelty and Future work.

Figure 1.2 illustrates the remainder of this thesis structure and outline.

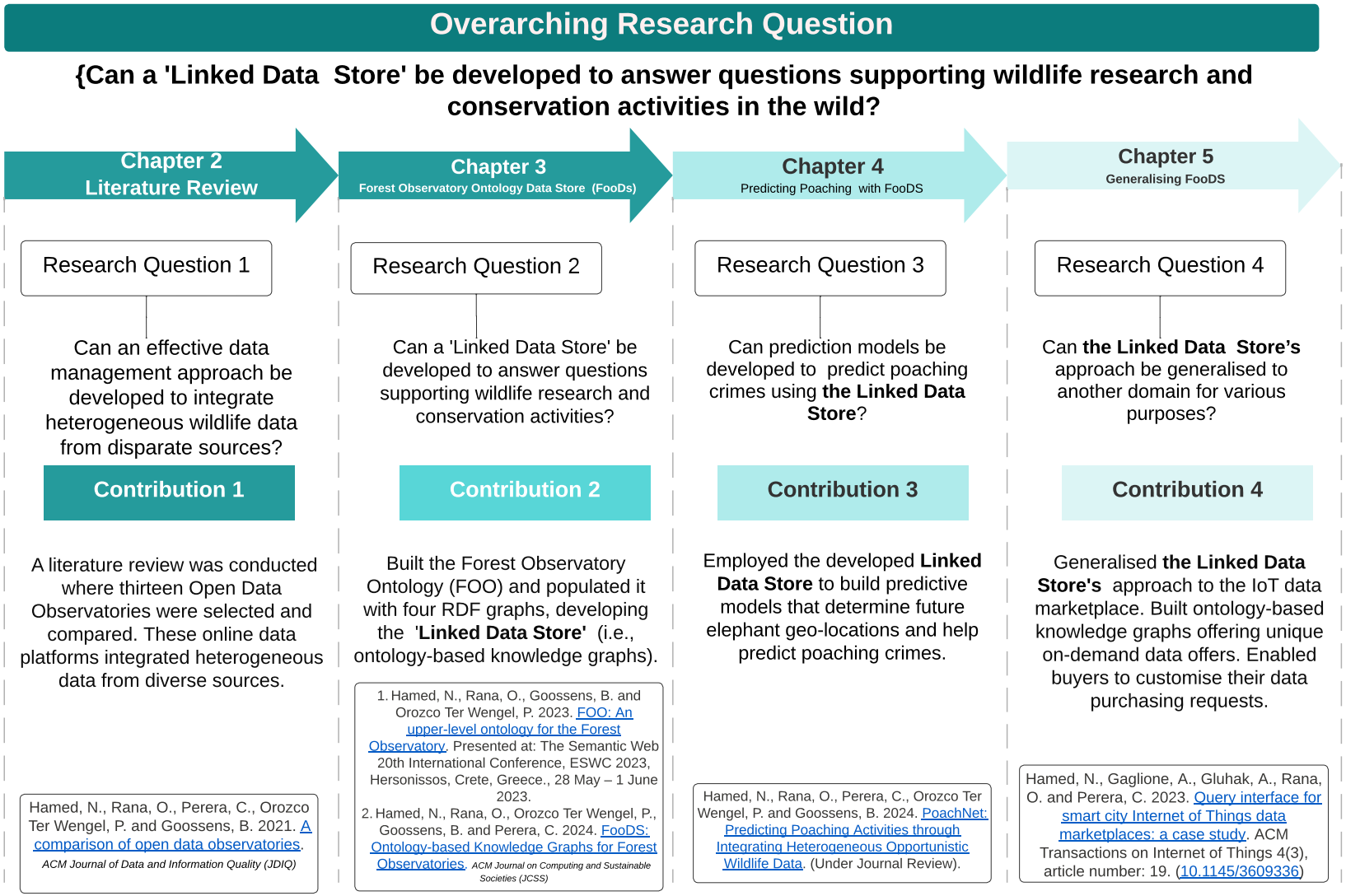


Figure 1.2 Thesis Structure

Chapter 2

Literature Review

This chapter addresses the first research question (RQ1): *Can an effective data management approach be developed to integrate heterogeneous wildlife data from disparate sources?*. Starting by investigating data types shared as Open Data to understand the roles of data management approaches. Following that, research techniques assisted in selecting Open Data Observatories (i.e., data platforms or data environments that integrate data from disparate sources to service a specific domain). The selected Open Data Observatories were examined in terms of their data themes, data management approaches, strengths and limitations, and possible research challenges faced whilst building them. The purpose of this literature review is to identify the most suitable data management approach to address RQ1. Once the suitable approach is identified, the chapter reviews relevant research on wildlife data integration and delves into key aspects of semantic web technologies, such as ontologies and knowledge graphs, emphasising their role in integrating wildlife data and predicting crimes, particularly those involving wildlife.

2.1 Introduction

Structured, semi-structured, and unstructured data can be generated from diverse sources, including government authorities, academic institutions, and citizens. These data categories apply to every sort of data, with structured data including inventories and catalogs organized in tables, semi-structured data such as operational manuals in JSON (JavaScript Object Notation) and XML (eXtensible Markup Language) formats, and unstructured data including text and media. These data are collected through various methods, such as questionnaires, web scraping and Internet of Things (IoT) devices. Whilst many governments have embraced the "Open Data" principles and made some of their data public, some commercial organizations collect large volumes of data, but only a fraction is accessible. Open Data refer to data

that are made available to the public by governments, organizations, and individuals [195]. They promote transparency, collaboration, and innovation, which can improve the quality of scientific research and contribute to the development of a sustainable ecosystem [51, 151].

Open Data portals, Open Data Observatories, and Repositories represent distinct systems within the data-sharing ecosystem, each serving unique functions and targeting specific audiences. Open Data portals serve as gateways to a wide range of datasets and resources from various sources. They provide search and discovery tools, data visualisation capabilities, and options for downloading data [57]. Open Data portals are centralised platforms where governments, non-profit organizations, and private companies release datasets to the public, aimed at enhancing transparency, enabling societal and economic benefits, and fostering innovation through open access to information on a variety of topics such as government operations, demographics, and economics [161].

Open Data Observatories are online platforms that curate and integrate real-time and historical data from different sources, presenting them in a unified manner. They focus on monitoring and analysing specific datasets for trends and insights, typically in public or research domains. The reliance on Open Data Observatories has become increasingly crucial in tackling the complex challenges faced by contemporary society and the environment. Previous research initiatives in [8] developed methods to survey Open Data platforms, providing insights into their availability and helping data publishers select the most suitable platforms for their data. A series of studies by Miller et al. [176], Moustaka et al. [178], Ma et al. [164], and Liu et al. [159] provided an understanding of the role of Open Data platforms in areas such as urban sustainability, smart city analytics, and ocean science.

Repositories provide broad platforms for sharing diverse research outputs. They can be domain-specific (storing data from a specific subject or field) or Generalist (serving multiple domains). Stall et al. [235] introduced the Generalist Repository Comparison Chart (GRCC) to assist researchers in identifying a generalist repository when a domain-specific repository [110] is unavailable for storing their research data. Generalist repositories (e.g., Zenodo, Figshare, and Dryad) archive diverse types of scholarly work, including datasets, articles, and preprints, thus supporting interdisciplinary research and increasing the visibility and impact of academic work beyond traditional publication venues. Such repositories require users to deposit their research outputs under open licenses, ensuring accessibility for further use. Most of this chapter aims to compare different Open Data Observatories and highlight their distinct features, methodologies, and challenges, thereby addressing thirteen Open Data Observatories, their data management approaches, some of their strengths and limitations, and primary research challenges faced when building them.

Table 2.1 Description and comparison of Open Data principles proposed by Sebastopol (S), the Sunlight Foundation (SF), and how they map to the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles.

Principle	Description	S	SF	FAIR
1. Complete	Data must be a complete and accurate representation of the original observations, including all computational details.	*	*	Findable
2. Primary	Data collected at the source and with metadata.	*	*	Findable
3. Timely	Data published promptly after collection.	*	*	Accessible
4. Accessible	Data must be easily accessible both physically and electronically.	*	*	Accessible
5. Machine-processable	Data in a format that can be easily processed by computers.	*	*	Interoperable
6. Non-discriminatory	Data is accessible to anyone without restrictions.	*	*	Accessible
7. Non-proprietary	Data in a format that does not require proprietary software.	*	*	Interoperable
8. License	Data freely available without restrictions or with clear permitting licensing for reuse	*	*	Reusable
9. Permanence	Data remain accessible online, including all versions.		*	Accessible
10. Usage costs	Accessing and obtaining data incur no fees.		*	Accessible- Reusable

2.2 Open Data

Open Data are free digital data, typically shared under open licences and organised in structured formats that follow established and agreed-upon standards. Open data are often supplemented by metadata, which provides "data about the data", such as data provenance information and data dictionaries. Metadata helps users understand the content datasets. Open Data are also released in formats designed to be easily read and processed by computer applications [151], enabling automated analysis and integration [264].

2.2.1 Open Data Principles

The expansion of Open Data is influenced by fundamental frameworks such as the Berners-Lee Five-Star Model [195] principles. Berners-Lee Five-Star Model evaluated Open Data on a scale from one to five stars, with higher ratings indicating open, machine-readable data and compliance with open standards. Kucera et al. [149] investigated the challenges of publishing and reusing Open Government data, including establishing a publication methodology within the COMSODE project, highlighting the role of Open Government Data in fostering transparency and citizen engagement. Open Data principles, further expanded

upon by groups such as the Sebastopol [261] attendees and the Sunlight Foundation [88], establish a framework to verify that government data are openly accessible. The FAIR data principles [267, 130, 26] provide a set of guidelines aimed at enhancing data reusability for both humans and machines, focusing on data being *Findable, Accessible, Interoperable, and Reusable.*

Findable : For data to be findable, it must be easily discoverable by both humans and machines. Achieving this involves assigning datasets persistent identifiers, such as DOIs (Digital Object Identifiers), which provide a permanent link to the data. In addition, detailed and descriptive metadata is crucial, allowing potential users to understand the nature, scope, and relevance of the data. This metadata should be stored in well-established and searchable repositories, ensuring that researchers can locate the data using common search tools and databases.

Accessible: Beyond being discoverable, data must also be accessible to those who need them. This does not imply that all data must be open access, as there are legitimate cases where restrictions may apply—such as privacy concerns or proprietary information. However, even in cases of restricted access, the metadata should remain open and accessible, informing users about the data's existence and providing instructions on how access can be requested. Accessibility also depends on the use of standardised, well-documented protocols.

Interoperable: In modern research, the greatest insights often come from integrating datasets from diverse sources, enabling new avenues of analysis. For this reason, data must be interoperable—capable of being combined with other data and integrated into different platforms and tools. This requires the adoption of widely accepted formats and standards. Moreover, the use of common vocabularies, ontologies, and taxonomies is critical integrating data from different domains.

Reusable: The final principle, reusability, emphasises the long-term utility of data. For data to be reused by others—whether for replicating a study, conducting new analyses, or applying them in different contexts—they must be accompanied by thorough documentation. This includes details on how the data were collected, processed, and analysed, as well as any relevant limitations or uncertainties. Furthermore, data should be released under clear and appropriate licensing terms.

Table 2.1 integrates Open Data principles, as discussed by both the Sebastopol group and the Sunlight Foundation, with the broader framework of the FAIR data principles, providing a comparative overview of their alignment. It shows ten critical principles identified for the openness and availability of government data. Moreover, it introduces considerations for non-proprietary formats, licence freedom, permanence, and the waving of usage costs to foster a more inclusive and accessible digital ecosystem. This alignment is further enhanced

by indicating which of these Open Data principles correspond to which element of FAIR data principles.

2.2.2 Open Data Sources

Open Data can arrive from different sources, varying from country to sector. The primary Open Data providers in the United Kingdom (UK) and the United States of America (USA) are governmental agencies such as (data.gov.uk) and (data.gov). International organisations, including the World Bank (data.worldbank.org), provide global datasets for the health, environment, and education sectors, including governments, academic institutions, and citizens. Educational and Research Institutions such as Harvard Dataverse (dataverse.harvard.edu/) and PANGAEA (pangaea.de/) often share research findings with the public.

In many academic institutions, publishers increasingly require researchers to make the data contributing to a paper available (i.e., making their data available for others to use and build upon, including surveys and observational data that provide empirical evidence.) By sharing data free of charge, researchers can collaborate, replicate ideas in different domains, and expand scientific knowledge. In the past two decades, citizens generated a high volume of data from smartphones and wearable devices (smart watches) [144]. These generated data include information collected through social media platforms, GPS tracking devices, and mobile applications. Sensor networks also contribute data on environmental conditions, vehicle movement, and electricity consumption.

2.3 Research Method

The nominated and employed research method to select the Open Data Observatories is SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research type) [54]. SPIDER is a framework for conducting rigorous, transparent, and reproducible reviews. It was employed for its flexibility, which can cope with the evolving nature of technology compared to other research methods like snowballing, which rely on academic literature's provenance. To widen the search, keywords were extracted for each SPIDER component based on synonyms and related terms derived from the thesis research questions. Searches were achieved using the Google search engine, Google Scholar, ACM digital library, and Cardiff University library, focusing on the following terms:

1. **S**ample: Open Data observatory.

2. **P**henomenon of Interest: domain-specific and multi-domain data observatory.

3. **D**esign: Open Data platforms.
4. **E**valuation: relevance, transparency, accessible.
5. **R**esearch type: descriptive, survey, research article.

2.3.1 Search Plan

The search plan used the Boolean operators AND and OR to connect the search items corresponding to each SPIDER component. This approach constructed search queries that incorporated relevant terms. For instance, the search query for the SPIDER elements would look like this: Sample AND Phenomenon of Interest AND Design AND Evaluation AND Research type ("Open Data platform*" OR "Open Data observatory") AND ("domain-specific data observatories" OR "domain-specific observatory" OR "multi-domain observatory" OR "data integration") AND ("accessible online platforms" OR "data platform") AND ("relevance" OR "transparency" OR "rigour") AND ("descriptive" OR "survey").

Using the OR operator within parentheses, we expanded the search to include variations and synonyms for terms such as "Open Data platform" and "Open Data observatory." We incorporated terms related to the phenomenon of interest, such as "domain-specific data observatories," "domain-specific observatory," "multi-domain observatory," and "data integration." To capture different aspects of the design and evaluation, phrases like "accessible online platforms" and "data platform" were included. Moreover, terms related to the desired research attributes, such as "relevance," "transparency," and "rigour," and the research types, such as "descriptive" and "survey", were added. This search strategy ensured a comprehensive coverage of relevant literature and maximised the chances of identifying relevant studies.

2.3.2 Observatories Selection Process

SPIDER search plan yielded a vast number of online data platforms. To filter out the relevant Open Data Observatories, specific inclusion and exclusion criteria were set to refine the selection process and ensure that only the most relevant platforms were included in our study. Filtering criteria, as shown in Figure 2.1, were based on domain experts' suggestions, platforms' establishment date, and relevance to the overarching research questions. By setting the inclusion and exclusion criteria, the most recent platforms available in the English language were detected, prioritising platforms that demonstrated clear relevance to the overarching research question.

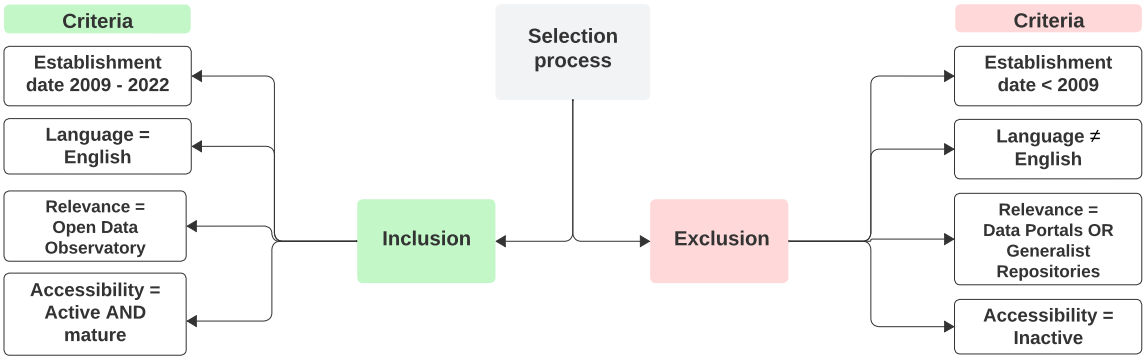


Figure 2.1 Inclusion and exclusion criteria for selecting the reviewed Open Data Observatories.

2.4 Open Data Observatories

The initial search process yielded forty Open Data environments. Each was manually checked to ensure it met the Open Data Observatories criteria. Through this evaluation, thirty-four Open Data Observatories were filtered out and identified. After a thorough manual evaluation, we arrived at a final selection of thirteen Open Data Observatories that satisfied all the necessary criteria. Therefore, the selected Open Data Observatories are introduced and discussed in the subsequent section- starting from the older ones and progressing to the newer ones (Figure 2.2). Each observatory is concisely outlined and characterised by its attributes, kinds of data, and significant accomplishments or obstacles.

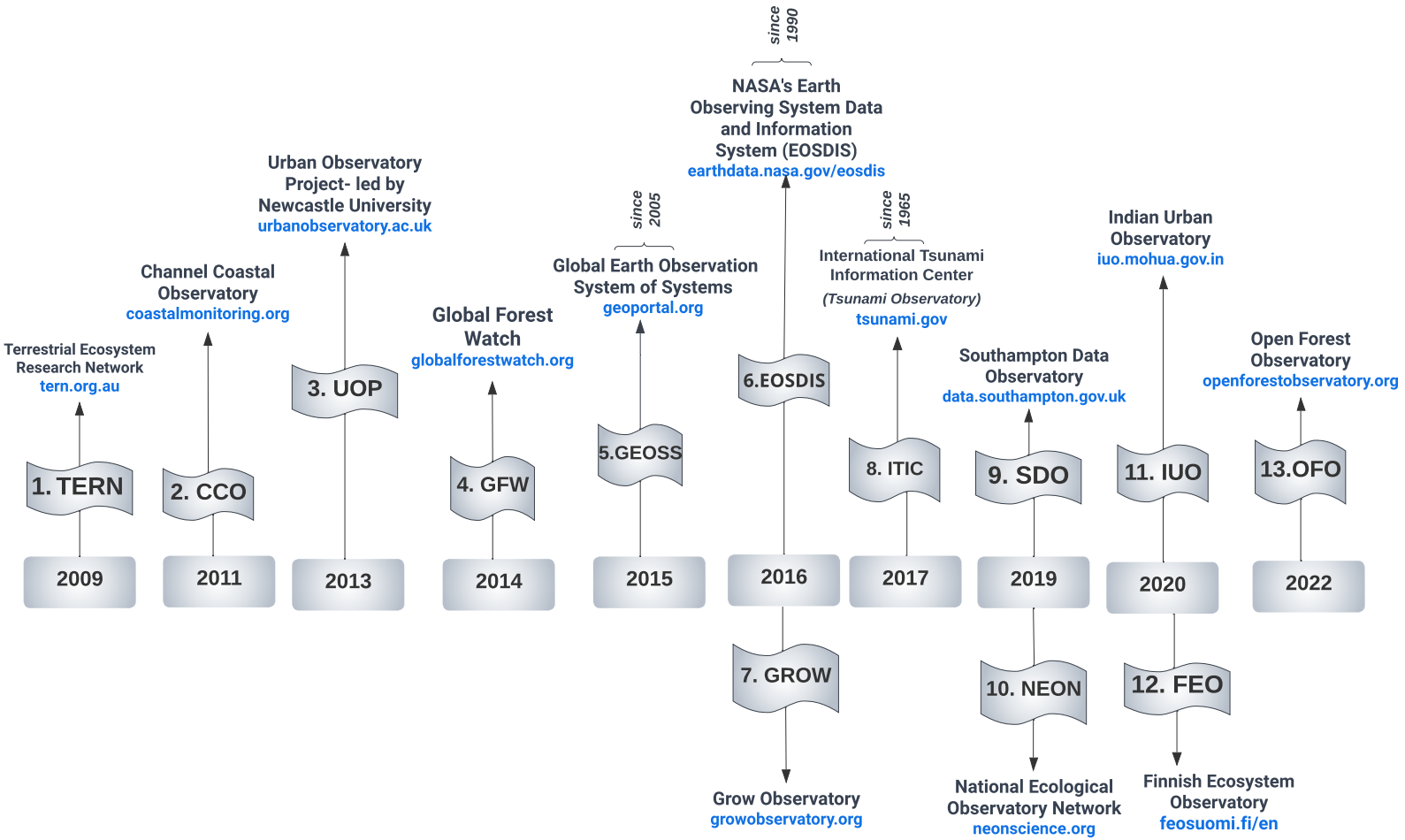


Figure 2.2 Timeline displays the selected Open Data Observatories. 1. Terrestrial Ecosystem Research Network (TERN) [50], 2. Channel Coastal Observatory (CCO), 3. Urban Observatory Project (UOP), 4. Global Forest Watch (GFW) [263], 5. Global Earth Observation System of Systems (GEOSS) [49, 59], 6. NASA's Earth Observing System Data and Information System (EOSDIS) [23], 7. Grow Observatory (GROW), 8. International Tsunami Information Center (ITIC)- Tsunami Observatory, 9. Southampton Data Observatory (SDO), 10. National Ecological Observatory Network (NEON) [21], 11. India Urban Observatory (IUO) 12. Finnish Ecosystem Observatory (FEO) [254], 13. Open Forest Observatory (OFO).

2.4.1 Terrestrial Ecosystem Research Network (TERN)

Terrestrial Ecosystem Research Network (TERN)[1] is a national research infrastructure programme in Australia that supports ecosystem science. The Australian Government established TERN in 2009 in response to a growing demand for a coordinated approach to terrestrial ecosystem research. The network, as such, comprises a range of field sites and data infrastructure that supports long-term environmental monitoring and scientific research, including biodiversity and land surface properties. TERN's infrastructure has over 600 ecological monitoring sites across Australia and advanced data management systems that allow researchers to access and analyse data from disparate sources. TERN aims to support evidence-based decision-making for ecosystem management and conservation in Australia and to promote a greater understanding of terrestrial ecosystems and their role in maintaining global environmental health.

TERN hosts a substantial and growing collection of diverse ecosystem datasets across Australia, covering topics such as vegetation, soil, and phenology. TERN provides a variety of data tools and services, including SHaRED for data submission and harmonisation, aligning with the FAIR principles, a Data Discovery Portal for accessing diverse ecosystem datasets, tools for data analysis and visualisation such as MCAS-S and the Data Visualiser, cloud-based research platforms like CoESRA, and resources for field data collection, including a network of monitoring sites. In addition, the Threatened Species Index- TSX (tsx.org.au) is a dynamic tool that helps understand how Australia's threatened species are faring over time. It provides visualisations and observations on temporal trends for 286 species of threatened and near-threatened mammals, birds, and plants in Australia.

2.4.2 Channel Coastal Observatory (CCO)

Since 2011, the National Network of Regional Coastal Monitoring Programmes has supported six projects along the English coastline. The overarching objective of these projects is to gather in-situ coastal monitoring data [154]. However, Contarinis et al. [53] highlighted some inconsistencies in the data quality and the methodologies employed. The Channel Coastal Observatory (CCO)[2] was established in response to these challenges. In England, 520,000 properties face the risk of coastal flooding, while 8,900 are threatened by coastal erosion. The CCO provides fit data for decision-makers in understanding coastal behaviour and identifying potential risks associated with coastal flooding and erosion [169]. The observatory covers various coastal regions (e.g., Northeast, East Riding of Yorkshire, Anglian, Southeast region

[1] tern.org.au/

[2] coastalmonitoring. org/

(low-lying land), and Northwest). The primary data types collected and displayed on its platform are topographic and hydrographic surveys. Topographic surveys entail beaches, cliffs, dunes, and coastal defence structures, whilst hydrographic surveys extend from the Mean Low Water (MLW) contour to 1 kilometre offshore. The CCO also generates heterogeneous real-time data on waves, tides, meteorology, and GPS measurements, which helps understand and manage coastal environments. Through its public API, developers can access and integrate the real-time coastal data (waves, tides, and meteorology) collected by the monitoring programmes. It also provides information on accessing coastal data through Web Map Services (WMS) in GIS software such as ArcMap and QGIS.

2.4.3 Urban Observatory Project (UOP)

The Urban Observatory Project (UOP)[3] was launched in 2013 and sponsored by the UK Collaboratorium for Research on Infrastructure and Cities (UKCRIC) - led by Newcastle University in collaboration with five other British universities; Sheffield, Bristol, Cranfield, Birmingham, and Manchester. The UOP monitors and analyses urban areas by deploying different sensors across UK cities. Real-time data from sensors and smart cameras allow the monitoring of urban dynamics. Each participating university focuses on specific aspects of urban life. For instance, Sheffield Urban Flows Observatory examines the impact of energy and resource flows on economic performance and social well-being. At the same time, Bristol Urban Flows Observatory transforms Bristol into a living laboratory for community engagement.

Cranfield Urban Observatory provides data-centric and remote-sensing solutions for addressing environmental, social, and economic issues. Birmingham Urban Observatory monitors critical infrastructure and its interplay with the environment, economy, and society. Lastly, Manchester Urban Observatory collects, analyses, and shares urban data to support informed decision-making at the city level. The collaboration between these observatories contributes to a better understanding of urban dynamics and assists efficient urban development [233]. The UOP's data types include traffic flow, parking spaces, cycling docking, pedestrian count, weather data, air quality, water quality, seismic activity, noise level, water-level (rainfall), beehives, energy usage data, thermal imaging, visual and hyper-spectral mapping, social media feeds, employee feedback, and quantifying the impacts of COVID-19 measures.

[3] urbanobservatory.ac.uk

2.4.4 Global Forest Watch (GFW)

Global Forest Watch (GFW) initiative[4] is a non-profit organisation that is part of the World Resources Institute (wri.org). GFW collaborates with over 100 organizations to provide a transparent and actionable platform supported by satellite technology and cloud computing. Such observatory empowers diverse stakeholders, wildlife practitioners, companies, and governments in forest management and combating deforestation. The GFW's web-based platform (observatory), launched in 2014, provides data and tools for monitoring forests and land use. The platform has amassed over four million users worldwide, benefiting diverse groups such as local law enforcement, park managers, international corporations, and civil society organisations in their endeavours to safeguard forests.

GFW's primary applications include the Forest Watcher mobile app for real-time threat detection, GFW Pro for managing deforestation risks in supply chains, and the Global Forest Review (GFR) for monitoring global forest objectives. Moreover, national governments employ GFW's technology for forest resource management, whilst small grants and fellowships support additional advocacy and research. Collectively, GFW assists in forest surveillance and management, combats illegal deforestation, promotes sustainable commodity sourcing, and supports conservation research on a global scale. GFW data types include satellite imagery for observing changes in forest cover, forest change data for tracking deforestation and regrowth, and land cover data for understanding land usage in addition to data about biodiversity, climate dynamics, and commodity supply chains, as well as legal and administrative boundaries, fire alerts, and water resources. GFW provides both developer-focused tools (APIs and open-source code) and a user-friendly MapBuilder platform to enable the creation of customised interactive mapping applications that leverage GFW's robust spatial data and analysis capabilities.

2.4.5 Global Earth Observation System of Systems (GEOSS)

Global Earth Observation System of Systems (GEOSS)[5] was created due to directives from the 2002 United Nations World Summit on Sustainable Development and the G8's 2005 commitment. Its purpose is to improve the development and application of earth observation technologies for environmental monitoring. Starting in 2005 with a 10-year implementation plan (2015), GEOSS aimed to provide coordinated, sustained observations of the Earth, focusing on nine key societal benefits (e.g., sustainable agriculture, biodiversity conservation, and climate change adaptation). The success of GEOSS's first decade led

[4] wri.org/initiatives/global-forest-watch

[5] geoportal.org/

to the implementation of a renewed 10-year plan (2016-2025), which aligned well with global initiatives such as the UN Committee of Experts on Global Geospatial Information Management (UN-GGIM) and the G8 Open Data Charter to enhance data sharing and management. GEOSS became a global partnership that advocated for the importance of Earth observations and collaborated with stakeholders to address global challenges. One of GEOSS's achievements was the establishment of its data-sharing principles, which advocated for Open Data access. These principles influenced standard data policies like the European Union's Copernicus programme [59]. GEOSS integrates heterogeneous data, aiming to facilitate continuous Earth system observations. Examples of data types are satellite imagery, atmospheric data, oceanographic data, geological data, and biodiversity information.

2.4.6 Earth Observing System Data and Information System (EOSDIS)

The Earth Observing System Data and Information System (EOSDIS)[6] is an active part of NASA's Earth Science Data Systems Program. The observatory integrates data from disparate sources, such as satellites, aircraft, field measurements, and other programs. EOSDIS supports Earth Observing System (EOS) satellite missions by handling command and control, scheduling, data capture, and initial processing tasks. These operations are overseen by NASA's Earth Science Mission Operations Project. EOSDIS's Science Operations, managed by NASA's Earth Science Data and Information System Project, entail generating higher-level science data products (levels 1-4), archiving, and distributing data products from EOS missions, in addition to other satellite missions, aircraft, and field measurement campaigns. This function is carried out within a distributed system that consists of interconnected nodes of Science Investigator-led Processing Systems and Distributed Active Archive Centres (DAACs), which are discipline-specific. EOSDIS offers a variety of curated data types that are crucial for evaluating ecosystem conditions, predicting species' geographical distributions, identifying materials based on spectral properties, and monitoring human-induced environmental changes. These data types include vegetation health, spectroscopy, species distribution, and ecological change tracking.

2.4.7 Grow Observatory (GROW)

Grow Observatory (GROW)[7] serves as a citizens' observatory, enabling individuals and communities to take proactive measures about soil and climate across Europe. GROW collected soil moisture, temperature, and light level data from low-cost "Flower Power" sensors

[6] earthdata.nasa.gov/eosdis
[7] growobservatory.org/

deployed across 24 locations in 13 European countries. This resulted in a 6,502 ground-based soil sensor network and 516 million rows of soil data datasets. Citizen scientists, community members, and land managers installed and maintained sensors voluntarily. Sensors' data were collected at 15-minute intervals and manually uploaded to the GROW servers using mobile phones. GROW integrated the sensors' data through a dedicated online hub, allowing members to register and visualise their sensors through time-series graphs and maps.

GROW also used GEOSS (observatory 5) data to provide public access to earth observation data collection. Further, data acquired from GEOSS were then used to more accurately predict extreme events, such as floods, droughts, and wildfires. In addition, GROW data played a significant role in validating and calibrating satellite observations, such as those from the European Space Agency's (ESA), SMOS (Soil Moisture and Ocean Salinity) mission and the future SMAP (Soil Moisture Active Passive) satellite. Artists and designers have played a significant role in GROW, with the former creating artworks reflecting the significance of soil ecosystems and remote sensing satellites and designing dynamic visualisations for agriculture and climate forecasting. It has also helped farmers in the Canary Islands reduce their water usage for irrigation by 30%. GROW received awards, including the Land and Soil Management Award 2019, the Stephen Fry Award for Excellence in Public Engagement 2020, and recognition as the first in the European Commission's annual GEO Plenary Statement on significant Earth Observation developments in 2019.

2.4.8 Tsunami Observatory

In March 2017, NOAA's National Tsunami Warning Center and Pacific Tsunami Warning Center, in partnership with the Tsunami Service Program, centralised their information on a Tsunami Observatory[8]. As a hub for information on tsunamis, it provides warnings, advisories, watches, and threat evaluations for Alaska, British Columbia, Washington, Oregon, and California regions. The observatory supplies real-time updates on event magnitude, depth, coordinates, and the time the seismic event occurred. It also shares bulletins and statements about the current tsunami status, clearly indicating if there are any warnings, advisories, watches, or threats in effect for the mentioned areas. *(*This tsunami observatory) educates civilians about tsunami risks following seismic activities, promoting safety and preparedness among residents of affected regions. It also communicates with other relevant observatories, such as the Deep-ocean Assessment and Reporting of Tsunamis (DART) project, a component of the U.S. National Tsunami Hazard Mitigation Program. DART employs seafloor bottom pressure recorders (BPR) and surface buoys to identify and report

[8] tsunami.gov

tsunamis in real-time. DART system has two generations, with the second-generation DART II enabling bidirectional communication since 2008. This system can detect tsunamis as small as 1 cm and transmits information to ground stations through a GOES satellite link for early detection and data collection. Moreover, the NOAA Tsunami Stations offer information on tide stations equipped to detect tsunamis along various coastlines. At the same time, the IOC Sea Level Monitoring Facility provides real-time monitoring of sea level stations worldwide.

2.4.9 Southampton Data Observatory (SDO)

Southampton Data Observatory[9] collects data from various stakeholders in Southampton and Hampshire and combines them with nationally published data, providing access to professionals, businesses, the voluntary sector, citizens, and communities. The observatory has been developed in partnership with statutory partners, including the National Health Service (NHS) Hampshire, Southampton, and Isle of Wight (CCG), and Southampton Voluntary Services, with data contributions from other partners such as the National Office of Statistics (ONS), Hampshire Constabulary, Hampshire Fire and Rescue Service, and South Central Ambulance Service. SDO is accountable to the Southampton Health and Well-being Board and the Southampton Safe City Partnership for delivering the Joint Strategic Needs Assessment (JSNA) and the Safe City Strategic Assessment. It considers data protection issues and ensures sufficient safeguards and disclosure controls are in place to protect the identity of individuals. SDO's data types include links to demographics, economy, education, health, housing, road safety and environment specific to Southampton and its immediate surroundings within the United Kingdom.

2.4.10 National Ecological Observatory Network (NEON)

The National Ecological Observatory Network (NEON)[10] is an Open Data observatory funded by the National Science Foundation. Initiating its operational phase in the summer of 2019, NEON allows access to data on various topics, including climate, land use, and biodiversity. NEON adopts a specialised method for selecting its study locations across the United States, including Hawaii and Puerto Rico, to capture diverse environmental conditions. The areas are split into 20 distinct zones, each comprising its own set of ecosystems, landscapes, and natural processes. As a result, NEON integrates extensive data on the well-being of plants and animals, soil and water quality, and many more. It uses state-of-the-art sensor technology

[9] data.southampton.gov.uk/
[10] data.neonscience.org/

and direct field observations. Notably, NEON collects and provides standardised data on a continental scale collected from 81 field sites equipped with automated sensor systems. NEON's focus on long-term, standardised data collection enables researchers to track and analyse changes in ecological systems over time to understand the impacts of climate change and other environmental factors. NEON engages scientific communities by encouraging researchers to use the available data in their research projects.

2.4.11 India Urban Observatory (IUO)

The India Urban Observatory (IOU)[11] is an Open Data Observatory established by the Ministry of Housing and Urban Affairs (MoHUA). IOU is a central hub for data and analytical tools related to the country's urban areas to equip policymakers, researchers, and citizens with reliable information on urban planning and development. It provides evidence-based decision-making and improves urban planning, offering city-level indicators of population statistics, infrastructure development, and economic growth information. Furthermore, IOU also provides data about water supply, sanitation, and waste management. Visualisation and analysis tools are available at the IOU to enhance data reuse and understanding. These tools enable users to explore and interpret the data in a user-friendly manner, aiding informed decision-making.

2.4.12 The Finnish Ecosystem Observatory (FEO)

The Finnish Ecosystem Observatory (FEO)[12] is a research and monitoring platform that serves as a resource for obtaining high-quality ecosystem data across diverse terrestrial and aquatic ecosystems in Finland. FEO allows researchers, policymakers, and the general public access to data and observations. Available data include climate, hydrology, biogeochemistry, and biodiversity. FEO employs eddy covariance flux towers, radiometers, anemometers, and infrared gas analysers to gather the required data. FEO provides standardised field monitoring methods, calibration guidelines, and field data collection apps to ensure consistent and reliable data collection. Mäyrä et al. [170] combined deep learning and remote sensing to enhance forest monitoring by classifying tree species using airborne hyperspectral imagery and LIDAR data. The study conducted in Finland's Boreal forests demonstrated the effectiveness of high-resolution hyperspectral data and ground reference measurements in efficiently distinguishing between different tree species for improved biodiversity monitoring.

[11] iuo.mohua.gov.in/portal/apps/sites
[12] feosuomi.fi/en/

2.4.13 The Open Forest Observatory (OFO)

The Open Forest Observatory (OFO)[13] employs drones and Artificial Intelligence (AI) to map and identify trees without needing traditional ground surveys. It establishes more than 100 forest plots, each roughly 25 hectares in size, to gather data vital for forest management in the face of issues such as droughts and wildfires. This initiative aims to improve forest and disturbance ecology research by creating three innovative cyberinfrastructure tools. The first tool is an AI-driven software workflow that transforms drone-captured imagery into forest inventory information– creating maps that accurately locate and scan individual trees. The second tool is an open database that contains tree maps from over 100 plots, each covering 25 hectares. These plots are managed with existing forest inventory networks (NSF's NEON) and cover a range of environmental and disturbance gradients. Lastly, the initiative includes documentation and training programmes, both online and in-person, to empower researchers to generate and share their data. This observatory applies photogrammetry to create 3D models of forest structures.

Moreover, it uses computer vision methods, supported by neural networks, for accurate species classification and to filter out incorrect tree identifications. The National Science Foundation primarily funds the OFO with additional support from The Nature Conservancy. The OFO is housed in three academic institutions: the Department of Plant Sciences at the University of California, Davis, the CIRES Earth Lab at the University of Colorado, Boulder, and the Bio5 Institute at the University of Arizona. It relies on ground-reference forest inventory data from the USDA Forest Service Pacific Southwest Region and the National Ecological Observatory Network (NEON) 2.4.10. The OFO also uses CyVerse and Jetstream2 computing infrastructure to support its operations.

2.5 Data Themes, Sources, Processing and Visualisation

This section discusses the data from the selected Open Data Observatories, examining their themes, sources and methods employed in their processing. Our thematic analysis, referencing [39], revealed two main themes: urban and no-urban data. We started the thematic analysis by reading through the data types collected for the selected observatories and taking notes. Table 2.2 shows data types the chosen observatories manage. Then, using NVIVO 12 software, we generated codes that helped us with the data themes. Words coded under "Transport" indicate `urban data`, whilst the words coded under "Soil Data" and "Seismic Events" entail `non-urban` data themes.

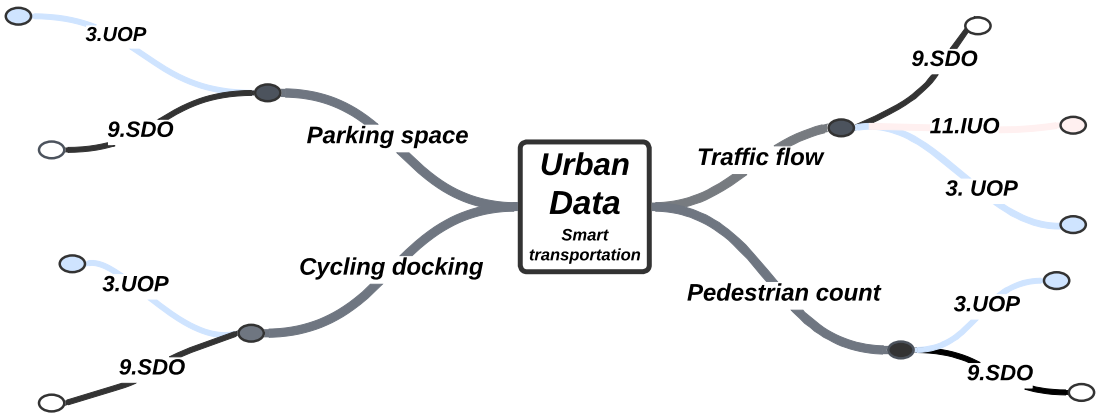
[13] openforestobservatory.org/

Table 2.2 Lists the Open Data Observatories and their data types.

Open Data Observatory	Data types
1. Terrestrial Ecosystem Research Network (TERN)	Vegetation, soil, and phenology.
2. Channel Coastal Observatory (CCO)	Topographic and hydrographic surveys. Real-time data about waves, tides, weather and GPS data.
3. Urban Observatory Project (UOP)	Urban data include traffic flow, parking spaces, cycling docking, pedestrian count, weather data, air quality, water quality, seismic activity, noise level, water level (rainfall, river and tides), beehives, energy usage data, thermal imaging, visual and hyper-spectral mapping, social media feeds, employee feedback.
4. Global Forest Watch (GFW)	Satellite imagery, biodiversity, soil, climate dynamics, commodity supply chains, legal and administrative boundaries, fire alerts, and water resources.
5. Global Earth Observation System of Systems (GEOSS)	Satellite imagery, soil, atmospheric data, oceanographic data, geological data, biodiversity information, and climate metrics.
6. Earth Observing System Data and Information System (EOSDIS)	Soil, vegetation, spectroscopy, species distribution, and environmental change.
7. Grow Observatory (GROW)	Soil, temperature, and light level.
8. International Tsunami Information Center (ITIC)	Water-level data, historical tsunami, recent tsunamis.
9. Southampton Data Observatory (SDO)	Urban data include links to demographics, economy, education, health, housing, road safety and environmental data.
10. National Ecological Observatory Network (NEON)	Soil, atmospheric data for climate change, biogeochemistry, ecohydrology, land cover processes, organisms, populations, and communities.
11. Indian Urban Observatory (IUO)	Urban data include population statistics, infrastructure development, economic growth, water supply, sanitation, and waste management.
12. Finnish Ecosystem Observatory (FEO)	Climate, soil, hydrology, biogeochemistry, and biodiversity.
13. Open Forest Observatory (OFO)	Forest drone imagery, forest structure metrics, tree sizes and species

2.5.1 Urban Data Theme

Urban data refer to information generated from activities taking place in cities, including data on smart transportation, human behaviour, and demographics. Smart transportation data are generated by devices (cameras) connected to the Internet of Things (IoT) and monitor traffic flow, vehicle counts, public transit usage, parking availability, congestion levels, average speeds, and pedestrian counts. Many observatories, the UOP, SDO and IUO, collect and analyse various types of urban data. Examining the UOP, it was noticed that it's focused on real-time data on city transportation (e.g., traffic congestion, parking availability, and public transit usage). On the other hand, SDO aggregates forwarding links from different stakeholders and Open Data providers (ONS) to data on transportation usage and behaviour, such as walking, cycling, and driving patterns, as well as transportation infrastructure (i.e., roads and public transit systems). IUO also collects data on transportation infrastructure (roads, highways, railways), usage, and behaviour (vehicle ownership, mode choice, travel patterns). These observatories have a common goal: *how urban transportation systems function and how they can be improved to better meet the needs of city residents*. The data collected by these observatories cover a range of urban data metrics, as analysed in Figure 2.3. Environmental data are collected in cities by one of the UOP observatories. To illustrate the concept, Figure 2.4 shows the environmental data types and parameter counts at Newcastle's



3. Urban Observatory Project (UOP) 9. Southampton Data Observatory (SDO) 11. India Urban Observatory (IUO)

Figure 2.3 Transport data metrics collected by Open Data Observatories.

Urban Observatory Project. Table 2.5 lists examples of the data types' parameters and their measuring units. Here, weather data refer to temperature, humidity, wind speed, and precipitation through a network of sensors deployed across Newcastle and the surrounding territories, and the water level data entail river and tide levels. Raw data were obtained from (newcastle.urbanobservatory.ac.uk/api-docs/doc/sensors-dash-types-csv/).

2.5.2 Non-urban Data Theme

Non-urban data refer to observations from areas outside city boundaries (i.e., rural, forest, and natural environments). Such data collected by our selected Open Data Observatories as listed in Table 2.2 span various environmental variables helpful in understanding ecosystem dynamics, climate change, and biodiversity. To mention a few observatories catering for non-urban data, TERN supplies data on vegetation, soil, and phenological data. The CCO delivers topographic, hydrographic, meteorological, and GPS data relevant to coastal dynamics. GFW and GEOSS use satellite imagery to monitor biodiversity, climate dynamics, and environmental changes. EOSDIS provides and analyses soil, vegetation, and ecological change data. GROW contributes data on soil conditions, temperature, and light levels. NEON offers comprehensive soil, atmosphere, biogeochemistry, and biodiversity data to track climate change impacts. The FEO and the OFO deal with data generated from boreal and temperature forests, respectively. The diversity of non-urban data from some of the selected observatories supports a holistic understanding of Earth's non-urban environments, facilitating research and conservation efforts across multiple disciplines.

Table 2.3 Compares Open Data Observatories, including their data types, geographic scope and the data themes they provide.

Observatory	Geographic Scope	Data Types	Urban Data	Non-Urban Data
1. TERN	Australia	Mangroves, vegetation, soil, phenology	No	Yes
2. CCO	UK	Topographic and hydrographic surveys, real-time coastal data	No	Yes
3. UOP	UK	Urban data (traffic, air quality, noise, water level, etc.)	Yes	No
4. GFW	Worldwide	Satellite imagery, biodiversity, climate dynamics, fire alerts	No	Yes
5. GEOSS	Worldwide	Satellite imagery, atmospheric, oceanographic, geological, biodiversity, climate metrics	No	Yes
6. EOSDIS	USA	Soil, vegetation, spectroscopy, species distribution, environmental change	No	Yes
7. GROW	Europe	Soil, temperature, light level	No	Yes
8. ITIC	Worldwide	Water-level data, historical tsunami, recent tsunamis	No	Yes
9. SDO	UK	Demographics, economy, education, health, housing, road safety, environment	Yes	No
10. NEON	North America	Soil, atmospheric data, climate change, biogeochemistry, ecohydrology, land cover processes	No	Yes
11. IUO	India	Population statistics, infrastructure, economic growth, urban services (water, sanitation, waste)	Yes	No
12. FEO	Finland	Climate, soil, hydrology, biogeochemistry, biodiversity	No	Yes
13. OFO	USA	Forest drone imagery, forest structure metrics, tree sizes and species	No	Yes

2.5.3 Data Themes Comparison

Table 2.3 provides a detailed comparison of the thirteen Open Data Observatories, each contributing to various domains of environmental and urban data monitoring. The observatories are evaluated based on their geographic scope, types of data collected, focus on urban or non-urban data, availability of data APIs, and common themes they address. For example, TERN in Australia is dedicated to environmental monitoring, including data on mangroves, vegetation, soil, and phenology. On the other hand, the UOP in the UK focuses on urban dynamics by collecting data on traffic, air quality, noise, and water levels. Observatories like NASA's EOSDIS offer extensive APIs for data access, facilitating broader research applications, while initiatives such as GROW in Europe emphasise citizen science and soil data without providing an API.

Several observatories share common themes in their data types and focus areas (see Table 2.4). For instance, TERN, FEO, and NEON all concentrate on environmental monitoring and ecosystem research. Similarly, the UOP, SDO, IUO focus on urban dynamics, planning, and infrastructure. In contrast, observatories like the CCO and the ITIC are unique in their focus on coastal data management and tsunami monitoring, respectively. GFW and the OFO both emphasise forest management, although GFW provides global coverage using satellite imagery, the OFO uses drone imagery specific to the USA. NASA's EOSDIS and GEOSS

Table 2.4 Lists common themes and differences among Open Data Observatories.

Observatory	Themes	Common with	Differences
TERN	Non-urban	FEO, NEON, GEOSS, EOSDIS	Focuses on mangroves, vegetation, soil, and phenology specific to Australia
CCO	Non-urban	ITIC	Specialises in topographic and hydrographic surveys of UK coastal regions
UOP	Urban	SDO, IUO	Emphasises real-time urban data collection in the UK
GFW	Non-urban	OFO	Platform supported by satellite technology and cloud computing
GEOSS	Non-urban	TERN, EOSDIS, FEO, NEON	Comprehensive global data including atmospheric, oceanographic, and geological data
EOSDIS	Non-urban	TERN, GEOSS, FEO, NEON	Specific to NASA's satellite data, focuses on soil, vegetation, and environmental change
GROW	Non-urban	TERN	Focuses on citizen science in Europe, does not offer a data API
ITIC	Non-urban	CCO	Centralises tsunami-related data on a global scale
SDO	Urban	UOP, IUO	Aggregate various urban data including demographics and health data
NEON	Non-urban	TERN, FEO, GEOSS, EOSDIS	Offers comprehensive ecological data for North America
IUO	Urban	SDO, UOP	Focuses on urban services in India, such as water and waste management
FEO	Non-urban	TERN, NEON, GEOSS, EOSDIS	Specific to Finland's diverse ecosystems
OFO	Non-urban	GFW	Uses drone imagery focusing on USA

both engage in earth observation but differ in their specific data types and scope. GROW is distinct for its focus on citizen science and soil data in Europe.

2.5.4 Data Sources

Open Data Observatories obtain data from Open Data portals, wireless sensor networks, and smart devices. Wireless Sensor Networks (WSNs) play a significant role in *urban and non-urban* data collection [96]. A notable example is the UOP, which uses a network of over 3600 sensors to capture diverse data streams from different physical environments. GROW, as such, employs Flower Power sensors to monitor in-situ soil moisture, fertiliser levels, and air temperature at 15-minute intervals [147, 268]. Other technologies contributing data to these observatories include LIDAR, ARGUS cameras, and satellites. ITIC- tsunami observatory provides data on water levels and historical and recent tsunamis. The water-levels data sourced from the DART (Deep-ocean Assessment and Reporting of Tsunamis) system and the National Oceanic and Atmospheric Administration (NOAA) coastal water-level stations. The DART system obtains water-level data from bottom pressure recorders on the seafloor, which measure water pressure with a resolution of approximately 1 mm of seawater and take 15-second averaged samples. The data are then transmitted to a ground station via satellite telecommunications, enabling real-time reporting. The DART II systems transmit standard mode data containing 24 estimated sea-level height observations at 15-minute intervals once

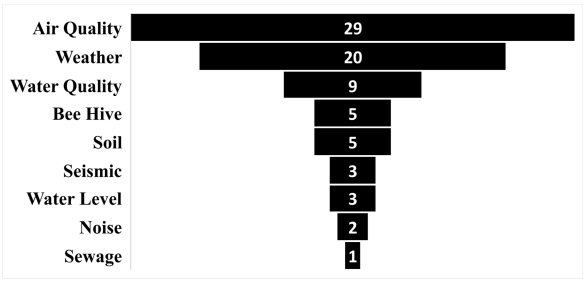


Figure 2.4 Newcastle Urban Observatory parameters count by data type.

Theme	Parameter	Unit
Air Quality	CO	ugm -3
Weather	Rain	mm
Water Quality	Dissolved Oxygen	mg/l
Bee hive	Brood nest temperature	Celsius
Soil	Soil Moisture	%VWC
Seismic	Vertical Displacement	m
Water Level	River Level	m
Noise	Sound	db
Sewage	Sewage Level	mm

Table 2.5 Newcastle Urban Observatory parameters examples and their measuring unit.

every six hours. The OFO uses drone imagery in a multi-step process to source data. First, numerous overlapping drone photos are taken from various angles to estimate each tree's three-dimensional structure.

Next, the Canopy Height Model (CHM) is generated by curating the data to create a high-resolution Digital Surface Model (DSM) that displays the vegetation's height in each pixel above the ground. Then, an algorithm identifies individual trees in the forest area using drone imagery and CHM data, resulting in tree-level maps of forest stands. NEON sources data and samples using a combination of automated instruments, field technicians, and airborne remote sensing. TERN gathers data using a variety of sensors, including eddy covariance flux towers, heat flux plates, radiometres, anemometres, infrared gas analysers, spectrometres, CosmOz soil moisture meters, groundwater bores, ecoacoustic sensors, phenocams, terrestrial laser scanners, UAV/drones, camera traps, and photopoints [226]. Table 2.6 groups and compares some of the observatories' primary data sources.

2.5.5 Data Processing

Most of the selected Open Data Observatories develop open-source software to ingest, harmonise, and integrate diverse data. Such data processing techniques are set to realise the potential value of Open Data by making them FAIR (Findable, Accessible, Interoperable, and Reusable) for researchers, decision-makers, and the broader community. TERN includes several tools and applications for data processing and analysis. To mention a few, SHaRED Data Submission (shared.tern.org.au) allows ecologists to upload their research data to the Australian Ecological Knowledge and Observation System (ÆKOS). It assists in creating structured metadata and assigns Digital Object Identifiers (DOIs).

Table 2.6 Lists and compares the Open Data Observatories data sources.

Open Data Observatory	Wireless Sensors	Smart Devices	Citizen Data	Weather Stations	Digital Cameras	Satellite/Lidar	Field Surveys	Sensing Vehicles	Drones	Crowd-sourcing
1. Terrestrial Ecosystem Research Network (TERN)	*	*		*	*	*	*		*	
2. Channel Coastal Observatory (CCO)	*	*		*	*	*	*	*	*	
3. Urban Observatory Project (UOP)	*	*	*	*	*	*		*	*	*
4.Global Forest Watch (GFW)			*			*	*			*
5. Global Earth Observation System of Systems (GEOSS)	*	*		*		*	*		*	
6. Earth Observing System Data and Information System (EOSDIS)	*	*		*		*				
7. Grow Observatory (GROW)	*	*	*	*			*			*
8. International Tsunami Information Center (ITIC)	*	*		*		*				*
9. Southampton Data Observatory (SDO)		*	*							*
10. National Ecological Observatory Network (NEON)	*	*	*	*	*	*	*	*	*	*
11. India Urban Observatory (IUO)	*	*	*							*
12. Finnish Ecosystem Observatory (FEO)	*	*				*	*			
13. Open Forest Observatory (OFO)			*						*	*

CoESRA Virtual Desktop (coesra.tern.org.au) enables access to a web-based virtual desktop from any device and is equipped with scientific software such as RStudio, Jupyter Notebook, and QGIS. EcoImages (ecoimages.tern.org.au) is a repository that organises images of vegetation, soil, and landscapes. To process live streams of diverse data, the UOP deploys real-time machine learning models on CCTV feeds and uses data queues, data sharding, and many edge processors along with hourly replication to reduce the occurrence of problems during live data streaming. To produce forest datasets, GFW uses machine learning to detect and map tree cover and loss, involving image segmentation, classification, and change detection. At the ITIC tsunami observatory, raw data from the tide gauges and DART buoys are processed by the PMEL (Pacific Marine Environmental Laboratory) and NGDC (National Geophysical Data Center) to remove errors and archive. NEON developed proprietary software to process raw data from sensors and field apps into standardised data products. NEON employs a unique "NEON Ingest Conversion Language" to establish and update data processing protocols as necessary. The OFO presents three cyber-infrastructure innovations to enhance data processing capabilities. These include a scalable, reproducible, AI-enabled software workflow for converting drone imagery into forest inventory data, a searchable database of treemaps that are aligned with forest inventory plot networks and accessible to the public, and documentation and training resources to encourage researchers to contribute their research data and analytical tools. Moreover, research [277] offers resources for individuals who want to create efficient and detailed tree maps of conifer forests without requiring extensive customisation of image acquisition and processing parameters.

2.5.6 Data Visualisation

Data visualisation transforms raw data into meaningful graphical representations that intended audiences can readily perceive, read and understand [273]. The selected observatories employ various visualisation tools and methods to present and communicate their collected data. Some of the data visualisations entail static and interactive maps [82], charts such as time series, scatter plots, histograms [234], bar, and pie graphs. For instance, TERN-ANU Landscape Data Visualizer (maps.tern.org.au) is a user-friendly atlas that offers spatial data on Australian landscapes, soil, ecosystems, and water resources. The data can be visualised on a map and explored through time-series data for specific locations. The UOP has interactive maps, digital comparison tools, thematic cartography, and real-time data visualisations to explore and monitor urban dynamics.

NEON collaborates with Google to enhance the visualisation and accessibility of its environmental data via the Google Cloud Platform, incorporating tools such as Google Earth Engine and BigQuery. This integration enables users to engage with and visualise extensive NEON datasets directly in the cloud. Global Forest Watch (GFW) visualises data through its Open Data portal, interactive map features, downloadable datasets, geospatial monitoring frameworks, and software like the Forest Trends Analysis Tool. EOSDIS visualises data through the Earthdata Cloud, providing users free access to NASA Earth science data for research purposes. ITIC - tsunami observatory offers real-time and historical tsunami data through 1-minute water level readings, event search tools, and interactive maps. IUO illustrates data stories and interactive maps using ArcGIS, thematic dashboards, and an Open Data portal to share urban vision with its designated stakeholders. GROW uses interactive maps and visualisation tools to visualise the collected actual soil moisture data and share them with its network [236].

2.6 Research Challenges

Establishing Open Data Observatories involves addressing various challenges related to integrating diverse data sources and systems. These challenges include ensuring data interoperability, scalability, and replicability since each data source has its own design and computing specifications. Combining and merging disparate data, without careful consideration, can lead to service conflicts, resulting in degraded data quality, loss of data provenance, and potential privacy breaches. This section explores these challenges, as depicted in Figure 2.5 and how each observatory addresses each challenge.

2.6.1 Data Integration

Data integration is the process of combining data from disparate sources into a unified view. Integrating heterogeneous data can positively impact decision-making; however, achieving valid integration faces many challenges, as noted by various researchers [20, 66]. Figure 2.5 outlines the main data integration components that Open Data Observatories may encounter. The interoperability challenge refers to the difficulty of integrating and harmonising disparate data sources and systems. Interoperability is one of the Open Data FAIR principles, as explained in section 2.2.1 [188, 20]. Integrating data from disparate sources may also involve using ontologies, managing large data volumes, and handling high-velocity data streams. Effective use of APIs is crucial for accessing and integrating data from different platforms.

To overcome this challenge, several observatories implemented various strategies. For instance, TERN harmonised the plot-based ecology using EcoPlots (ecoplots.tern.org.au), a *semantic data integration* system that maps each data source to *TERN Plot ontology*. The term 'ontology' is a structured framework that defines the relationships between concepts within a specific domain, providing a shared vocabulary for that domain [111]. OWL (Web Ontology Language) is a formal language used to create and share these ontologies on the web, enabling better data interoperability. The UOP deployed a platform called the "Urban Data Exchange (UDX)" (urbandatacollective.com/urban-observatories-case-study) that acts as a central hub for onboarding, harmonising, and serving the real-time data streams from the different urban observatory systems. EOSDIS enhanced data interoperability through standardisation of data formats and metadata, a distributed and interoperable architecture across nodes like the Science Investigator-led Processing Systems (SIPS) and Distributed Active Archive Centers (DAACs), which enabled efficient data retrieval [216].

2.6.2 Data Quality

Applied research defined the term data quality differently [207], a commonly used definition by Strong et al. [237] describing data quality as data fit for the intended purpose. Byabazaire et al. [38] and Taleb et al. [244] testified that data quality is a mature research topic in big data and database management. However, Perez-Castillo et al. [207] claimed its youth in Smart Connected Products (SCP) [281] and the Internet of Things. Data quality plays a significant role in Open Data Observatories, as a sufficient quality level can build trust between the cyber and physical world [38, 207].

Each observatory addresses data quality using different strategies, the UOP manages data quality by using automated checks for data anomalies, calibrating sensors against precision stations, and incorporating user feedback. The UOP also acknowledges the limitations

of low-cost sensors and design their data use accordingly. GFW ensures data are up-to-date by automating updates or requesting providers to notify them of changes. EOSDIS methodology ensures metadata quality of Earth observation data hinges on a framework prioritising correctness, completeness, and consistency. NASA uses automated and manual reviews to identify and rectify issues, demanding active collaboration with data providers to implement enhancements [37]. The CCO and NEON implement quality assurance and control practices. The CCO ensures the reliability of marine observations, flagging poor data but not eliminating them, whilst NEON applies rigorous quality measures to ensure data quality. For example, observation system data use mobile apps with constraints and validation rules. Instrument System data benefit from sensor placement, maintenance, and calibration.

Airborne Remote Sensing data are calibrated and tested pre- and post-flight. Automated checks and expert reviews ensure reliability, while flags and metrics provide transparency. IUO handles quality through trusted data sources, accuracy, transparency, and interactive visualisations but has limitations in completeness and update frequency. The OFO prioritises data quality through standardised, open-source workflows for drone-based forest mapping, accessible via its GitHub repository. It also employs cloud-based tools to process drone imagery into detailed forest maps, facilitating ease of use as well as a central database to support data sharing and quality enhancement through community feedback. As shown in Figure 2.5, data quality challenges in the selected Open Data Observatories are closely related to the FAIR principles, particularly data findability, accessibility, and reusability. Using trusted sources and maintaining rigorous data entry standards minimise anomalies. Sensor calibration, data entry rules and constraints implementation provide reliable data and enhance their accessibility. Data completeness and consistency through quality assurance processes also contribute to better metadata and documentation, making the data reusable.

2.6.3 Data Provenance

Data provenance, which traces the origins and lineage of data, is crucial in Open Data Observatories. Maintaining rigorous data provenance allows observatories to ensure data transparency, reliability, and reproducibility [9, 202, 124]. TERN releases weather data accompanied by their lineage, including (a) the type and model of the automatic weather station used for collection; (b) the specific location and characteristics of the site; (c) the instruments used for measuring different weather parameters, along with their accuracy and resolution; (d) the methodology for data recording and the intervals at which data were stored; (e) the procedures followed in case of sensor failure including using alternative data

sources for gap filling and indicating this within the dataset; and (f) the availability of the data and contact information for access to more granular data (hourly data).

Similarly, SDO commits to full metadata inclusion for all its published data compendiums and resources, encompassing data sources and time frames. NEON's dedication to rich metadata and thorough documentation strengthens the provenance and traceability of its data offerings. This commitment includes the provision of Digital Object Identifiers (DOIs) for NEON data packages, enhancing their findability and citability. NEON's approach to data provenance involves metadata management, adherence to FAIR principles, data citation tracking, and handling data from diverse sources, focusing on transparency and accessibility. In a different vein, research [265] recommends applying blockchain technology for data provenance.

Blockchain can revolutionise how data are managed, enhancing transparency, security, and trust. By leveraging its immutable ledger, data integrity and authenticity can be guaranteed, ensuring that once data are recorded, it cannot be altered. Moreover, the decentralisation offered by blockchain reduces risks associated with centralised data storage by distributing data across a network, thus enhancing data resilience and accessibility through peer-to-peer sharing. Furthermore, blockchain's encryption and smart contracts safeguard sensitive data and automate data access permissions, ensuring only authorised access. It also offers a transparent audit trail for all data modifications and transactions, facilitating traceable data lineage and enforcing open data licenses automatically. Data provenance in our selected Open Data Observatories aligns with the FAIR principles through elements like data access licenses, documentation, transparency, data lineage, and citations. As shown in Figure 2.5, clear data access licenses enhance accessibility and reuse, whilst documentation and transparency improve findability and interoperability. Data lineage ensures reliability and supports reusability, and citations facilitate proper attribution, enhancing findability.

2.6.4 Data Privacy

Data privacy is critical in protecting personal and sensitive information from unauthorised access and disclosure. Open Data Observatories implemented various measures to address data privacy challenges, including data anonymisation, access controls, and encryption [155, 168, 230, 218, 124]. These observatories handle massive amounts of data from various data sources through orderly collection, aggregation, and analytics. However, these data may contain sensitive details such as personally identifiable information and endangered species locations [230, 206, 4, 150, 71, 160].

TERN, the CCO, and the UOP all have dedicated privacy statements that outline their data privacy practices. These include compliance with regulations like GDPR, providing privacy

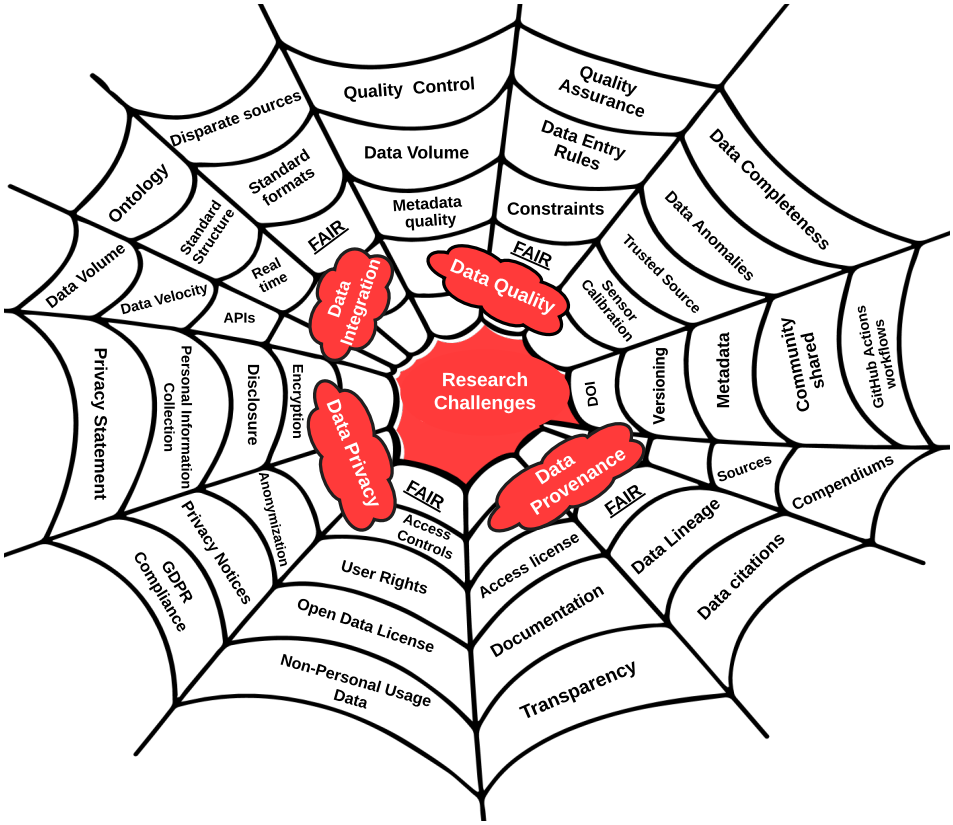


Figure 2.5 This diagram captures the intricate web of research challenges in data management, segmented into four primary categories, Data Integration, Data Quality, Data Privacy, and Data Provenance. Each challenge extends into related subtopics and approaches to overcome them that touch the periphery of the web, symbolising the complex and interconnected nature of these issues. The visual metaphor of a spider web conveys the idea that each aspect is a critical thread in the overall structure of data management.

notices, defining lawful data processing, implementing security measures, and respecting user rights. Similarly, GFW and GEOSS approach data privacy through transparency, consent-based processing, security, and clear points of contact for users. NASA's EOSDIS also has a privacy policy that emphasises protection and proper use of information in line with relevant laws and regulations. GROW addresses privacy by using an open data license, collecting only anonymised sensor data without personal identifiers, and operating under institutional oversight. The ITIC-tsunami observatory's privacy policy covers aspects like cookies, email handling, and user rights under the Privacy Act. SDO adheres to the overall privacy policy of Southampton City Council, whilst NEON securely manages user accounts, anonymises data reporting, and applies Creative Commons licensing. In contrast, IUO has a privacy-focused approach, avoiding automatic capture of personal information and only collecting such data if explicitly provided by users, with appropriate security measures.

Finally, the OFO focuses on openly sharing its forest mapping data and tools, rather than collecting or managing personal user information, implying a commitment to data transparency and accessibility. Data privacy in our selected Open Data Observatories involved encryption, access controls, disclosure of information, anonymisation, privacy notices, secure collection of personal information, privacy statements, and GDPR compliance. As shown in Figure 2.5, encryption and access controls ensure secure and restricted data access, aligning with FAIR principles. Disclosure and privacy notices enhance transparency, improving findability and interoperability. Anonymisation and secure collection practices ensure data reusability without compromising privacy. Privacy statements and GDPR compliance maintain legal and ethical standards, supporting data integrity and user trust.

2.6.5 Takeaways

The evaluation of Open Data Observatories, summarised in Table 2.7, underscores their strengths, including high-quality environmental monitoring data (TERN), tools for analysing coastal changes (CCO), and real-time monitoring of urban areas (UOP) and forests (GFW). These observatories also provide platforms for citizen engagement (SDO, GROW) and global data archiving (EOSDIS). Despite these strengths, challenges persist, such as limited geographic coverage (UOP, FEO), concerns over data security and quality (CCO, ITIC), inconsistent updates (IOU), and restricted data diversity (OFO). To address these issues, the recommendations focus on enhancing data quality and transparency (TERN, GFW, ITIC), expanding geographic reach (SDO, FEO), implementing real-time alert systems (CCO, EOSDIS), and adopting advanced technologies such as AI, blockchain, and drones (TERN, OFO). *Importantly, the semantic web emerges as a promising data management approach for overcoming these challenges. By leveraging ontologies, harmonised standards, and*

FAIR-aligned practices, Semantic web solutions enable interoperable, high-quality, and secure data integration across diverse observatory systems, as illustrated in Figure 2.5. This approach effectively addresses issues related to data coherence, provenance, and privacy, making it particularly suited to integrating wildlife data silos. As a result, Semantic web technologies qualify as the effective data management solution I have been searching for. The subsequent sections explore wildlife data management approaches, with a focus on semantic web applications for wildlife data integration and crime prediction, particularly in the context of wildlife crimes.

2.7 Data Management Approach for Wildlife Data

This section provides an overview of the relevant research on wildlife data management. It begins by outlining data integration in wildlife applications, then transitions to exploring how semantic web technologies are used to represent wildlife data and evaluates various methods. The section then discusses ontologies and knowledge graphs, explaining how they can be developed and why using ontologies to create knowledge graphs could benefit relevant stakeholders.

2.7.1 Data Integration for Wildlife Applications

In the last few years, data integration across wildlife applications has been investigated on multiple fronts to improve the collection, analysis, and use of wildlife data. For example, drones with AI are transforming conservation monitoring by providing vital information about wildlife numbers and distribution [104, 36]. Likewise, the Internet of Things (IoT) technologies have revolutionised wildlife tracking and provided an unparalleled opportunity for combined data capture and integration [158].

However, developments in machine learning techniques like Convolutional Neural Networks (CNN) have been used to classify wildlife and combine it with ecological data in a very cost-effective way [166]. Integrative models that combine information from extensive surveys and participatory science [179] are aiding in the understanding of wildlife population dynamics at broad geographic scales as well as over time. For instance, hierarchical models are being developed to integrate data from both count and distance sampling for improved wildlife population estimates [99]. Furthermore, incorporating Unmanned Aerial Vehicle (UAV) systems in combination with machine learning algorithms enables near-real-time wildlife monitoring and conservation [214]. Increasingly, novel machine learning applications are being used to decode animal movement patterns and improve our understanding of

Table 2.7 Strengths and limitations of the selected Open Data Observatories, future recommendations and some takeaways.

Data Observatory	Strengths	Limitations	Future Recommendation	Takeaways
1. TERN[14]	High-quality data on environmental monitoring, along with tools and expertise, provided to researchers.	Limited coherent national capability for monitoring freshwater ecosystems.	Integrating blockchain for data provenance and artificial intelligence for Linked Data.	Semantic data integration and the Threatened Species Index (TSX)[15]
2. CCO[16]	Access to tools and models to analyze coastal data and predict morphological changes.	Outsourcing data storage may impose security concerns.	Incorporate extreme events alert system.	Extreme events analysis.
3. UOP[17]	Ability to provide a wide variety of real-time and historical data on different aspects of the urban environment.	Urban observatories do not extend their coverage to all cities across the UK, resulting in a limited geographical reach.	Lack of evident research documenting the positive impact of the project (e.g., reduce crime rates).	Real-time data integration.
4. GFW[18]	Forest Watcher mobile app for real-time threat detection, the GFW Pro for managing deforestation risks in supply chains, grants and fellowships.	Limited data lineage.	Provide details how data are collected and evolved over time to enhance data provenance.	Real-time forest monitoring via satellite imagery and remote sensing.
5. GEOSS[19]	Data platform flexibility enabling users to adapt it to their needs.	GEOSS does not guarantee its Earth Observations' accuracy or take responsibility for their use.	Invest in quality assurance and control.	Platform flexibility.
6. EOSDIS[20]	Global, long-term and reliable Open Data.	Limited validation for satellite-based data with ground-based measurements.	Consider real-time update and alert system for extreme events.	Data long-term archiving useful for analysis and training AI applications.
7. GROW[21]	Empowers citizens and communities to have a say on soil and climate matters across Europe.	Limited data types.	Integrate more data sources such as air quality and noise level.	Citizen science.
8. ITIC[22]	Centralized and authoritative source for providing real-time information, and warnings about tsunami events and risks.	Data quality and provenance challenges causing errors in tsunami database.	Addressing data quality for improving the reliability and usability of the tsunami data.	Hazard alert system.
9. SDO[23]	Crowd-sourcing, allowing citizens to understand local issues and contribute to problem-solving in urban development and sustainability matters.	Lack of real-time data and APIs.	Extend geographic scope.	Civic engagement and transparency.
10. NEON[24]	Open Data with good quality and sufficient documentation.	Sensor locations at certain sites are seasonally adjusted or removed due to unfavourable or unsuitable measurement conditions.	Implement hybrid power solutions combining wind power, solar power and energy storage systems for the Oksrukuyik Creek (OKSR) site, where operations cease during winter.	Educational resources such as the learning and code hub.
11. IOU[25]	Wide range of urban data.	Inconsistent data frequency.	Consider using applications for data quality assurance.	Urban data diversity.
12. FEO[26]	Ongoing monitoring and research initiatives related to Finland ecosystems.	Limited data coverage, lack of data privacy statement.	Expand geographic scope.	Platform presentation in multiple languages.
13. OFO[27]	Educational resources to understand forests.	Limited data diversity, privacy policy not shared in the website.	Integrate more remote sensing wildlife data, supplemented with contextual information	Drones and Artificial Intelligence (AI).

species behaviour [187]. Cross-species toxicokinetic modelling Novel models are being used to assess the risk due to endocrine-disrupting chemicals in wildlife [269].

2.7.2 Semantic Modelling for Wildlife Data

Semantic web technologies enable data interoperability and integration of multiple types of wildlife data, leading to the development of knowledge graphs for querying and analysis [253, 90, 209]. Technologies like GPS tracking, Wireless Sensor Networks (WSN) [276, 224], and the devices connected to the Internet of Things (IoT) [34, 145, 232, 73] gather a lot of information about the environment and need ways to combine it, like knowledge graphs [272, 107, 196]. These graphs provide information about how species are connected, how ecosystems work, and how the environment affects wildlife [204, 14]. This helps people from different fields work together to protect species and make informed decisions [189, 194].

Previous research, such as Athanasiadis et al. [15] developed a semantic framework for significant carnivore conservation in northern Greece, integrating animal tracking data with ecological niche modelling for habitat suitability. In contrast, this work differs in location, integration process, and output flexibility and employs an executable interface for interactive analysis across various data types. Wang et al. [259] applied semantic technology to model wildlife observations, including pollution effects on ecosystems and storing provenance data for traceability. The semantic modelling in this study is similar in some ways, but it is different in how the data are transformed. Instead of Wang et al.'s manual RDF model conversion, this study uses the Resource Description Framework (RDF) Mapping Language (RML) and modular pipelines for scalable data conversion into triple data stores.

Researchers Mireku et al. [177] and Zheng et al. [225] use semantic inference to help find new information and make predictions about how animals will move in the future. On the other hand, Wannous et al. [262] worked on creating a trajectory ontology that includes parts of three main ideas: (i) moving objects, (ii) marine environments, and (iii) spatiotemporal models. This was accomplished by converting their data into an OWL ontology using an open-source tool (`uml2owl`). Wannous et al. constructed a domain ontology that integrates various sub-ontologies tailored to specific use cases. This research adopted a different approach that allowed semantically modelled data to populate a wildlife monitoring ontology. Furthermore, this method includes ontology documentation, publication, and maintenance plans as recommended features.

2.7.3 Wildlife Ontologies

In computer science, *ontology* is a formal and explicit specification of a conceptualisation used to represent knowledge in a particular domain [111]. Ontologies have been used in various domains, including biodiversity, to model knowledge [6]. Previously, the development of ontologies was based on manual curation by domain experts. However, this process is time-consuming and prone to errors. In the context of biodiversity, ontologies have been developed to represent concepts such as species, habitats, and ecosystems. The semantic web for Earth and Environmental Terminology (SWEET) [217] is an example of a large-scale ontology that covers several domains related to the environment.

The Wildlife Ontology (WO) [215] is another example of an ontology developed specifically for wildlife data. In principle, ontologies are logically well-defined vocabularies that link various data sources and define their connections firmly. They comprise classes, relations, and instances. Data entities are represented as graphs with nodes and edges using a data model like the RDF. Using the RDF model, a piece of information is converted into a graph composed of *(subject, predicate, object)*, for instance (Soil_ID, Soil_pH, 4.88). Ontologies can be expressed as a tuple of five elements [245], formulated as follows:

$$Ontology = (\boxed{\text{C}}, \boxed{\text{HC}}, \boxed{\text{R}}, \boxed{\text{HR}}, \boxed{\text{I}}) \tag{2.1}$$

Where:

C = (instances of "rdf:Class") stands for concepts.
HC = (“rdfs:subClassOf”) stands for concept hierarchy.
R = (instances of “rdf:Property”) stands for relationships between concepts.
HR = (“rdfs:subPropertyOf”) stands for relationship hierarchy.
I = (“rdf:type”) the instantiation of the concepts in a particular domain.

2.7.4 Ontology Development Methodologies

For a suitable methodology, I searched the ACM digital library (dl.acm.org) and Google Scholar (scholar.google.com). Our search terms included "ontology methodology," "ontology development methodology," and "ontology building approaches." Methodologies include the eXtreme design (XD) methodology [30], which is a modular, incremental approach that maps a set of competency questions to one or more Ontology Design Patterns (ODPs) [92] before integrating them into the ontology under construction.

The DILIGENT methodology [210] provides a more flexible trial-and-error approach, recommending the order of discussion, evaluation, justification, and testing in a use-case.

28

Table 2.8 Compares ontology development methodologies. CQs= Competency Questions, NLs= Natural Language Statements. Lightweight= Conceptualisation

Ontologies Development Methodologies	Reference	CQs	NLS	Tabular	Integration	lightweight	Formalisation	Implementation	Evaluation	Documentation	Publication	Maintenance
The eXtreme Design (XD)	[30]	*			*	*	*	*	*			
DILIGENT	[210]		*		*	*	*	*	*			
METHONTOLOGY	[162]	*			*	*	*	*	*			
On-To-Knowledge Methodology	[242]	*			*	*	*	*	*			*
Ontology Development 101	[190]	*			*	*	*	*	*			
NeOn Methodology	[238]	*			*	*	*	*	*			
Linked Open Terms (LOD)	[211]	*	*	*	*	*	*	*	*	*	*	*

METHONTOLOGY methodology [87], on the other hand, proposes a waterfall, an incremental development approach that focuses on the lightweight ontology version. Although METHONTOLOGY provides detailed guidelines for the life-cycle development of ontologies, it must be generalised to fit multiple domains. The On-To-Knowledge Methodology (OTKM) [242] focuses on the initial setup, enterprise applications, and maintenance of ontologies. Other well-known methodologies include "Ontology Development 101" by Noy et al. [190] and NeON by Suárez-Figueroa et al. [238]. Whereas the former focuses on ontology conceptualisation, the latter divides the ontology development process into nine distinct scenarios to accommodate a broader range of use cases.

Further ontology development methodologies were reviewed by Aminu et al. [12] and Singh et al. [228]. The Linked Open Terms (LOT) project [211] builds on over two decades of ontological engineering experience, taking inspiration from the Neon methodology [101]. It emphasises borrowing and reusing classes from related ontologies and allows for including natural language statements and tabular data during the requirement-gathering phase.

Moreover, LOT promotes the sharing of ontologies following the Linked Data and FAIR principles for the semantic web [28, 89] to facilitate their reuse by the research community and software applications. Table 2.8 compares the different ontology development methodologies.

2.8 Knowledge Graphs

A knowledge graph [114, 29] organises information into a graph structure, where nodes represent entities and edges define their relationships. The term "Knowledge Graph" gained

popularity with Google's Knowledge Graph project [75]. Following that, many academics have evaluated the term and used it in various contexts [201, 282, 46, 122, 45]. A commonly accepted definition of a knowledge graph captures knowledge by defining entities and their relationships [76]. Knowledge graphs offer several benefits in wildlife data management, enabling data integration, standardisation, linking, and reuse by combining characteristics of different data management paradigms.

2.8.1 Knowledge Graphs Creation Methodologies

A knowledge graph [114, 29] organises information into a graph structure, where nodes represent entities and edges define their relationships. Different methodologies exist for creating knowledge graphs, and the choice of method depends on factors such as the stakeholders involved, domain, intended applications, and available data sources. Some approaches include starting with an essential core and gradually enhancing it, following an Agile or "pay-as-you-go' approach [16]. Another strategy involves initiating a knowledge graph without predefining its schema (i.e., ontology) and gradually building schema and instances during creation. However, designing a knowledge graph schema beforehand can significantly enhance its utility [141]. A six-step process involving data identification, ontology construction, knowledge extraction, data processing, data integration, and knowledge graph evaluation is also commonly used [100].

Furthermore, employing robust tools for linked data, data integration, and data management whilst continuously analysing and adjusting deliverables is another viable methodology [22]. The ad hoc creation of knowledge graphs that reuse existing knowledge by interlinking relevant classes and properties from existing ontologies has also been practised [134]. The World Wide Web Consortium (W3C) (w3.org) recommends using RDF mapping languages (w3.org/TR/r2rml/), such as RML (rml.io/specs/rml/), R2RML (w3.org/TR/r2rml/), and xR2RML [175] for scalability and interoperability. RML is designed to map heterogeneous data structures onto the RDF (w3.org/RDF/).

The process starts by generating a text file defining the mapping rules that an RML processor executes to create the output RDF dataset [65]. Prior academic studies have extensively explored the development of semantic knowledge graphs and the evaluation of mapping languages and systems to generate RDF knowledge graphs from heterogeneous (semi-)structured data. Ryen et al. [223] and Van Assche et al. [250] contributed to the study area. In addition, Corcho et al. [55] presented a notable case in which they designed an ontology to create a knowledge graph for an ICT firm. These studies collectively emphasize the significance of semantic knowledge graphs and the utility of RDF-based approaches in representing and integrating data across various domains.

2.8.2 Ontologies for Knowledge Graphs

Using ontologies in knowledge graphs reduces ambiguity, ensures data compatibility, and establishes a formal representation of concepts and relationships [127, 148, 32]. With a defined ontology, data collection schemas from different sources can leverage shared vocabulary, resulting in semantic data integration. Ontology-powered knowledge graphs improve data interoperability, promote reusability and data exchange [136], enable automated reasoning, and enhance analytical capabilities. Table 2.9 compares the benefits of building a knowledge graph with and without an ontology.

Table 2.9 Compares the benefits of building a knowledge graph with ontology and without ontology

Benefits	With Ontology	Without Ontology
Less Ambiguity	Ensures a normalised representation of concepts and relationships.	Increased ambiguity in data and lack of a normalised structure.
Data Integration	Accelerates Data Integration.	Slower and more complex heterogeneous data sources.
Knowledge Representation	Enables complex relationship modelling and nuanced insights.	Limited ability to model relationships and capture intricate connections.
Data Interoperability	Facilitates seamless data exchange and system interoperability.	Challenges in integrating data from diverse systems.
Reusability	Promote ontology reuse and extension across applications and domains.	Lack of ontology reuse and extension leads to redundancy and inconsistency.
Reasoning	Enables automated reasoning and inference based on ontology relationships.	Limited ability for automated reasoning and logical inference.
Improved Search	Enhanced targeted search and querying through structured data representation.	Less precise and effective search and querying due to lack of structure.

2.8.3 Knowledge Graphs for Data Modelling

Knowledge graphs (KGs) have witnessed significant advancements in research, particularly in augmenting the capabilities of predictive models. Pahuja et al. [197] discussed the complexities of building prediction models with knowledge graphs, aiming to deduce new facts from existing data. Their research identified inherent challenges in traditional Graph Neural Networks (GNNs), such as over-smoothing and scalability issues. Addressing these, they proposed a novel "retrieve-and-read" framework, anchored by a Transformer-based GNN, which markedly improved the model's ability to extract and used relevant contextual information for predictions. Simultaneously, Duan and Chiang [69] introduced an integrated system to streamline complex predictive tasks for domain experts. This system, employing an

ontology-centric approach, consolidates diverse data into RDF knowledge graphs, facilitating efficient querying and analysis. A notable application of their system was demonstrated in a case study focusing on forecasting the future trajectory of fuel cell technologies. The system's proficiency in assimilating data from various sources, such as academic research and patents, showcased its versatility and effectiveness in predictive modelling.

In urban planning and analysis, Ning [186] introduced the Unified Urban Knowledge Graph (UUKG) dataset to overcome limitations in existing UrbanKGs. The project involved constructing comprehensive UrbanKGs for two significant cities containing millions of data triplets. The study unveiled intricate high-order structural patterns within these UrbanKGs through qualitative and quantitative analyses. The primary focus was enhancing urban *spatio-temporal predictions* by testing various KG embedding methods and integrating them into advanced spatiotemporal models. The UUKG dataset and its source code, made publicly available, represent a significant contribution to the field, encouraging further exploration and research.

Yan [271] explored the application of virtual knowledge graphs in the industrial sector, particularly in predictive analytics for hydraulic systems. This approach, aligned with the emerging needs of Industry 4.0, particularly in predictive maintenance, presents a pioneering method in digital modelling and system analysis. In the healthcare and pharmaceutical domains, Feng et al. [85] unveiled DKADE, an innovative framework combining deep learning with knowledge graphs to detect adverse drug events (ADEs). This framework addresses common issues in clinical narratives, such as missing drug information and the complexities of multiple medications. Complementing this, Zeng et al. [85] conducted an extensive review of KG methods in drug repurposing and adverse drug reaction predictions, focusing on the crucial role of knowledge graphs in drug discovery processes. In a related study, Wang [260] introduced KG-DTI, a deep-learning approach rooted in KGs aimed at predicting drug-target interactions for Alzheimer's disease treatments. This method uses a comprehensive KG containing thousands of positive drug-target pairs. It employs a Conv-Conv module to extract significant features, resulting in a highly effective neural network for drug-target interaction calculations. These diverse studies, covering urban planning, industrial maintenance, and medical research, exemplify the expanding influence and applicability of KGs. They highlight the potential of KGs to revolutionise predictive analytics in various fields, offering innovative solutions to complex challenges and paving the way for future developments in this dynamic and evolving area.

2.8.4 Knowledge Graphs for Crime Prediction

Tompson et al. [246] outline the integration of open data from various sources as a formidable approach to crime prediction. Knowledge graphs, in particular, are useful in organising, managing, and effectively using large volumes of information. Their ability to augment data with relationships and semantics transforms raw data into intelligent, explainable insights, as explored by Sikos [227]. To mention few work from the literature, Deepack and his team [63] addressed the recent surge in crime rates with a Bi-LSTM neural network tailored to classify a spectrum of crime types. Their approach involved data gathering from Google News and Twitter, including preprocessing, initial labelling via the Fuzzy c-means algorithm, vector creation using Term Frequency-Inverse Document Frequency, and feature extraction through GloVe word embeddings. The cornerstone of their model's enhanced classification capability was dynamically crafted ontologies derived from weighted graphs from news and social media sources. Wang and colleagues [258] developed HAGEN, an end-to-end graph convolutional recurrent network aimed at predicting various types of crime across different geographical areas. Their model incorporated a homophily-aware constraint, guiding the optimisation of the region graph to ensure adjacent nodes exhibited similar crime patterns, in line with diffusion convolution principles. Simultaneously, Iqbal and associates [129] used actual crime datasets from several U.S. states for crime category prediction. Their multifaceted approach included techniques such as data reduction, Naïve Bayesian methods, Decision Trees, and Confusion Matrices. Bogomolov et al. [31] sought to forecast crime hotspots in London, merging data from human mobile networks, demographic insights, and open crime data. They employed various techniques, from Logistic Regression to Neural Networks, Decision Trees, Random Forest, and K-fold cross-validation.

Almanie and team [10] explored real-world crime datasets for Denver and Los Angeles. Their methodology embraced Decision Trees, Naïve Bayesian analysis, and Confusion Matrices. Similarly, Chen et al. [44] harnessed data from Twitter alongside weather information to predict crime incidents, relying on linear modelling techniques. Kang et al. [137] introduced an innovative feature-level data fusion method, incorporating environmental context information to surmount the limitations of existing crime prediction models. They employed a deep neural network (DNN) that integrated diverse datasets, including crime statistics, demographic data, meteorological information, and imagery from Chicago, enriching the prediction model with extensive environmental insights.

2.8.5 Wildlife Crime prediction

In the realm of wildlife conservation and biodiversity, the development of significant databases and information systems, such as the Global Biodiversity Information Facility (GBIF) [152], has been useful. GBIF's role as an internationally-funded network collating data on Earth's life forms is vital for biodiversity records. Similarly, the Encyclopedia of Life (EOL) [200] augments these efforts by documenting species on the planet, integrating diverse information sources to provide comprehensive data on species taxonomy, distribution, and conservation status. Wikidata [255], with its open, editable nature, and the UK-based National Biodiversity Network (NBN) Atlas [212], especially for species and habitats in the UK, are crucial in making biodiversity data accessible. eBird [241], managed by the Cornell Lab of Ornithology, contributes significantly with its extensive database of bird observations, enhancing global understanding of bird distribution and abundance.

These resources become particularly evident in the context of wildlife crime, a pressing global issue that operates transnationally and requires an international mitigation strategy [180]. Wildlife trafficking criminals, including poachers, intermediaries, and end consumers, are often part of larger syndicates orchestrating illicit operations. These syndicates rely on the exploitation of biodiversity and are responsible for reimbursing poachers and couriers for obtaining animal parts. Research that addressed wildlife crime, Hofer et al. [121] analysed the economic facets of illegal hunting in the Serengeti. They modelled factors like weapon costs, hunting expenses, potential penalties, and income loss due to travel. They deduced poacher behaviours and motivations by employing herbivore population statistics and community hunting questionnaires. Bakana et al. [17] explored multimedia data mining techniques against poaching. They reviewed object detection, image classification, and behaviour analysis methods for poacher identification.

Haas et al. [115] emphasised social network analysis for targeting wildlife trafficking networks. The authors consolidated federated databases containing criminal intelligence data from heterogeneous sources. Usually, a member of the federation, who is the requester, sends an email containing a Structured Query Language (SQL) query to each federation member. When a federation member receives such an email query, they can disregard it if they don't trust the requester. Alternatively, they may execute the query on their local database and send the result back to the requester as an encrypted file attached to an email. As such, they equipped law enforcement with strategies to dismantle these networks, as evidenced by a rhino trafficking case study.

In 2018, Haas and colleagues developed a political-ecological simulator model to manage human-wildlife conflict in [116]. The model uses decision-making diagrams to represent groups and animals to describe ecosystems. The model's parameters can be statistically

fitted to a political-ecological action history dataset using consistency analysis. The fitted model can then be used to find politically feasible management plans to achieve conservation goals. The authors also developed a web-based system for automatically acquiring group action data to update the model's parameter values in real-time. Actual time was tested on the management challenge of conserving the South African rhino population in the face of severe poaching pressure. The model generated decisions that matched 82.4% of those in a real-world dataset. The authors concluded that anti-poaching can increase disincentives against incentives, but it may require a significant investment that is not socially acceptable. Providing employment can shift the trade-off at a given level of incentives for relatively low and politically feasible input.

Critchlow et al. [60] employed Bayesian hierarchical models on ranger patrol data in Uganda, identifying illegal activities and accounting for ranger observation errors. Their study discussed the impact of illegal resource use on biodiversity loss within protected areas. Data were collected by ranger patrols in the Queen Elizabeth Conservation Area (QECA), Uganda, to identify the patterns, trends, and distribution of illegal activities. Encroachment and poaching of non commercial animals were the most prevalent illegal activities within the QECA [248]. The study also found that ecological covariates were not valuable predictors for the occurrence of unlawful activities, but the *location* of illicit activities in previous years was more helpful. Regular patrols throughout the protected area, even in regions of low occurrence, are also required. Ranger patrol strategies must be implemented to target illegal activities, informed by the location of past occurrences of illicit activity, which is the most useful predictor of future events.

Previous research [275, 72, 143, 174, 219] stated that geospatial standards enhance data sharing, analysis, evidence-based decisions, and reduce wildlife trafficking. For this, Gore [108] stressed on the significance of *geospatial data standards in combating wildlife trafficking*. They employed participatory workshops, online platforms, and communication with over 100 participants globally to establish these standards, facilitating data-driven actions such as indictments and network disruption. As a result, an open-access wildlife trafficking data directory and visualisation tool for researchers were developed. Ferber et al. [86] examined wildlife trafficking's broader impacts, focussing on the challenges in its mitigation. They presented a data-driven model predicting trafficking routes.

Applying heuristics, Yang et al. (2014) launched the Protection Assistant for Wildlife Security (PAWS) to optimise ranger patrols using game theory. Nguyen et al. [185] built the so-called Comprehensive Anti-Poaching with Temporal and Observation Uncertainty Reasoning (CAPTURE). Used as an anti-poaching tool, Capture provided a noticeable improvement over the state-of-the-art [83, 121, 60] in many ways. CAPTURE accounted

for the rangers imperfect observations challenge, integrated the temporal effect on poachers behaviour, and overcame the requirement of knowing the poachers' numbers. Nguyen et al. introduced two heuristics (evaluation-and-feedback and trial-and-error methods): parameter separation and target abstraction, respectively.

However, CAPTURE's predictions had some ambiguity (too many targets), the learning process took a long time on a high-performance computing cluster, and the model learned was hard to interpret since it makes predictions based on linear combinations of different decisions. Kar et al.[139] investigated the limitation of CAPTURE [185] by presenting an adversary behaviour modelling system, INTERCEPT (Interpretable Classification Ensemble to Protect Threatened Species). Intercept used decision trees to make predictions that are easier to interpret and address the spatial correlation of the dataset by introducing a spatially aware binary decision tree algorithm (BoostIT). To further augment INTERCEPT's performance, they built an ensemble of the best classifiers, which boosted predictive performance by a factor of 3.5 over the existing CAPTURE model. Moreover, they provided an extensive empirical evaluation of the largest poaching datasets from Queen Elizabeth National Park (QENP) in Uganda. They analysed 41 different models and a total of 193 model variants and presented month-long field test results. One year later, Gholami et al. [98] combined capture & intercept models and presented a hybrid model predictive power with spatiotemporal power that selectively predicted snaring activity in precise locations in areas of high snaring activity, resulting in rangers finding more *snares and snared animals*. Their field test was much more prominent in scale than the other models, involving 27 patrol posts. Then, Gholami and McCarthy [98] developed an imperfect observation aWare Ensemble (iWare-E1) model to improve the detection of poaching crime in two protected areas in Uganda and Africa. The predictions were made on 14-year (2003–2016) datasets of type, location, and date of wildlife crime activities.

In contrast to these approaches, this work pioneers integrating heterogeneous wildlife data with deep learning on an ontology-based knowledge graph. This model achieves remarkable predictive performance by understanding complex *animal movements'* behaviour and constructing reasoning rules, outperforming conventional techniques and state-of-the-art methods. This innovative approach holds promise for advancing poaching prediction and enhancing wildlife conservation efforts.

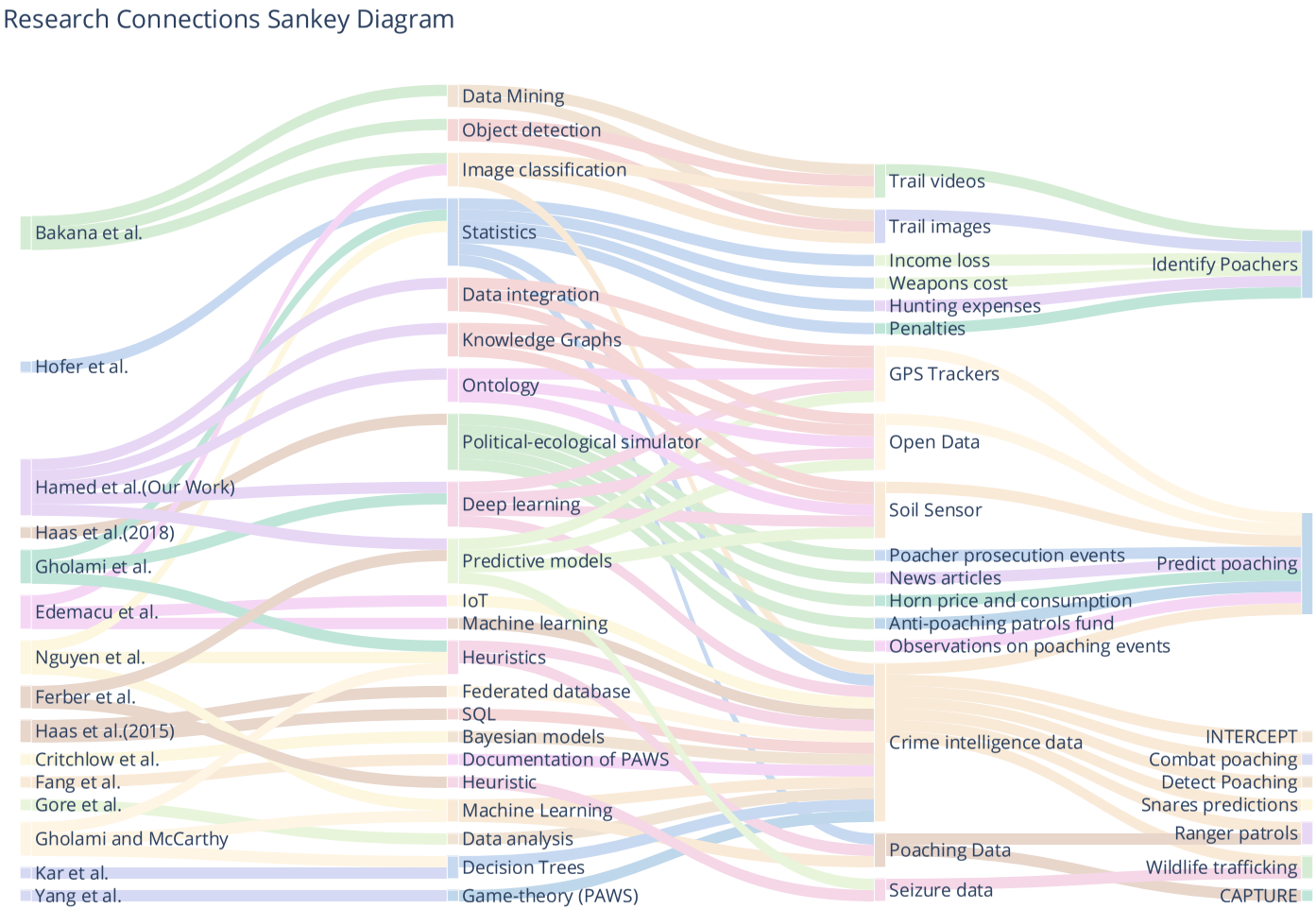


Figure 2.6 Multi-level Sankey Diagram to visualise wildlife crime prediction authors, their approach, data used and the resultant product. Bakana et al. [17], Hofer et al. [121], Haas et al. [115], Edemacu et al. [74], Ferber et al. [86], Haas et al.[116], Gore et al. [108], **Hamed et al.(this work)**, Critchlow et al. [60], Gholami et al. [97], Kar et al. [139], Yang et al. [274], Fang et al. [84], Nguyen et al. [185], Gholami and McCarthy [98]

2.9 Summary

This chapter introduced and analysed thirteen Open Data Observatories, offering data spanning both urban and non-urban settings on a global and regional scale. The main features, data availability, and usability of these observatories were examined. Despite challenges in comparing them due to varying sizes and stages of development, collaborations and connections were identified, such as between NEON and the OFO, and between GROW and GEOSS. In addition, the data explored in the Open Data Observatories were grouped into urban and non-urban themes, highlighting commonalities in data types and processing approaches across the observatories.

Research challenges related to integrating diverse data sources whilst maintaining their reliability and integrity were identified, including data integration, data quality, data provenance, and data privacy. Solutions to these challenges vary widely depending on the observatory's domain, data source, and target audiences. Specific strengths and weaknesses of each observatory were also pinpointed, forming the basis for future recommendations. These findings underline the importance of collaboration between different disciplines, the standardisation of data, and adaptable strategies to overcome data and system integration challenges. More specifically, the gap in the current literature that this research addresses is the lack of integrated platforms for wildlife data.

Through a comparative analysis of existing data management methods, semantic web technologies were identified as the most suitable data management approach for this research data, thereby addressing RQ1 and fulfilling C1. Semantic web technologies were selected for their unique ability to handle the integration of heterogeneous data sources whilst ensuring interoperability, standardisation, and scalability. In comparison, Geographic Information System (GIS) and Artificial Intelligence (AI) are highly effective for spatial analysis and predictive modelling but are often domain-specific and less adaptable for unifying diverse datasets across disciplines. Unlike semantic web technologies, GIS lacks built-in mechanisms for semantic integration, and whilst AI can process and analyse data, it typically requires pre-structured, high-quality inputs and lacks the capacity to natively enforce data standards or manage provenance. Semantic web, through the use of ontologies and knowledge graphs, excels at representing complex relationships enabling seamless integration of disparate wildlife data silos.

Consequently, an overview of the relevant research on wildlife data management examines how semantic web technologies are used to model wildlife data by comparing different approaches. Ontologies and knowledge graphs were discussed, focusing on their development and construction. In addition, the benefits of using ontologies to create knowledge graphs were justified. Further, previous studies applying knowledge graphs for data modelling, crime

prediction in general, and wildlife crime prediction in particular were reviewed. This review led to the identification of insufficient application of knowledge graphs in poaching prediction. This sets the stage for introducing an ontology created to standardise heterogeneous wildlife data, build wildlife knowledge graphs, and use them to predict poaching intents.

The proposed ontology, named the Forest Observatory Ontology (FOO), is the outcome of a collaboration between Cardiff University's School of Biosciences, the Danau Girang Field Centre (DGFC), and the School of Computer Science and Informatics. The following chapter details the ontology development lifecycle and the creation of its associated knowledge graphs, collectively referred to as the Forest Observatory Ontology Data Store (FOODS).

Chapter 3

Forest Observatory Ontology Data Store (FooDS)

This chapter addresses the second research question (RQ2): *Can a 'Linked Data Store' be developed to answer questions supporting wildlife research and conservation activities?* This chapter begins by briefly defining the term "Forest Observatory" and comparing various data management approaches in the context of Open Data Observatories. Following this, the chapter introduces an ontology named the Forest Observatory Ontology (FOO). Four semantically modelled wildlife datasets were used to populate FOO, resulting in an ontology-based knowledge graph named the Forest Observatory Ontology Data Store (FooDS), which serves as the 'Linked Data Store' to answer wildlife research questions. FOO and FooDS were evaluated using specialised open-source ontology scanners, feedback from domain experts, and applied use cases. This chapter contributes FooDS, the first ontology-based knowledge graph for Forest Observatories, which provides accurate query responses, reasoning capabilities, and granular data acquisition from diverse datasets. FOO in turtle format, FOO's documentation and FooDS in turtle format and their resource website are published at https://w3id.org/def/foo, https://w3id.org/def/fooDocs, https://w3id.org/def/fooDS, and https://ontology.forest-observatory.cardiff.ac.uk.

3.1 Introduction

Forest Observatories integrate and analyse wildlife data to answer questions that support data-driven analysis and forest monitoring [119]. Such observatories can enhance the understanding of ecosystems, species interactions, and environmental changes, aiding conservation efforts and informed decision-making [68]. In wildlife research activities, multiple methods

are employed to collect data, including field surveys, direct observation censuses, GPS tracking, motion-activated trail cameras and airborne sensors. However, the collected data often exist in silos or isolation due to the independent handling of maintenance, analysis, and storage by separate research activities. In addition, many environmental scientists lack expertise in managing data using computer science methods, which can lead to data management being overlooked rather than a planned process [222].

Siloed data hinder collaboration as groups work independently, thereby reducing opportunities for data sharing [167]. For example, consider one group studying the impact of elephant populations on soil health in a specific ecosystem, whereas another group investigating the behaviour and movement patterns of the same elephant population. The first group collected data on soil composition, nutrient levels, and erosion rates, whereas the second group collected information on migration routes, feeding habits, and social interactions. Soil researchers might need to understand elephant movement patterns to assess their impact on soil compaction and nutrient distribution, whereas elephant researchers could benefit from insights into how soil quality influences elephant grazing behaviour. Collaboration between these two groups can be facilitated by using a common *data store* that standardises the datasets and *links* their entities. This *linked data store* can integrate these diverse data in a way that is comprehensible to both humans and machines. Effective data management for Forest Observatories improves the long-term collection, quality, and persistence of data, enhancing the ability to address key ecological questions regarding conservation and natural resource management. Traditional data management strategies such as data warehousing and lakes are commonly employed to integrate data from various sources [243, 183, 266].

Data warehousing involves extracting, transforming and loading of data from different sources into a structured database system, ensuring uniform storage, and facilitating accessibility and analysis for data scientists. Conversely, data lakes serve as repositories for structured and unstructured data in raw formats. Conceptual models of how animals interact with and use habitats that link diverse research data also existed in past studies [123, 35, 40]. However, these approaches often lack meaningful connections between the data entities. Data scientists can derive substantial benefits from incorporating semantic webtechnologies such as ontologies and knowledge graphs into their workflow. The semantic web equips computers with the necessary tools and languages to understand and process the data in a way that is meaningful and useful for specific applications, enabling rule-based and automated reasoning, data integration, and complex querying capabilities.

Ontologies [111] are structured frameworks that describe the types, properties, and interrelationships of concepts within a specific domain. They serve as formal representations of a set of concepts and their connections, facilitating a shared understanding that can be

communicated between people and their computational systems. Knowledge graphs [118], on the other hand, represent a way of structuring and integrating knowledge based on relationships between entities (such as objects, individuals, concepts, or events), enabling machines and people to interpret and use interconnected information effectively. Ontology-based knowledge graphs focus on developing semantic relationships in data. These relationships form meaningful connections between concepts in a particular domain, enabling an understanding and interpretation of how these concepts relate to each other. The semantic web technologies enable precise querying, complex relationship analysis, semantic consistency, and data interoperability. Moreover, the reasoning capabilities can enable data scientists to infer implicit knowledge that is not overtly specified within the data.

Our research employed a novel ontology integrating elements from established ontologies to unify the Internet of Things (IoT) and wildlife concepts (biodiversity, conservation biology, habitat fragmentation, and endangered species management). We applied semantic modeling techniques to reformat various wildlife datasets into graphs and merged them with our ontology to produce four knowledge graphs.

This chapter's contributions to Forest Observatories include the following:

1. The Forest Observatory Ontology (FOO) and its knowledge graphs (FOODS), equipped with online documentation for describing wildlife data generated by sensors.

2. A resource website for FooDS, offering information on their creation and usage.

3. An analytical executable notebook and a dashboard to remotely query, visualise and analyse four distributed wildlife knowledge graphs in a granular, unified manner.

3.2 Forest Observatory Ontology (FOO)

The Forest Observatory Ontology (FOO) is proposed, a novel ontology that represents wildlife data collected through remote sensing devices. FOO articulates complex relationships and facilitates the linkage of diverse concepts through a versatile approach incorporating classes and properties from established ontologies. FOO standardises data entities and formalises their semantics, enabling the integration of diverse wildlife datasets from various sources. Specifically, it can articulate the relationship between an animal, a sensor, and its geolocation, and the observations collected when a sensor is attached to an animal record its geolocation. Furthermore, FOO facilitates the*semantic linkage of data sources* that share common concepts, thereby allowing for efficient retrieval of animal location data through sensor queries. In addition, FOO enhances data analysis capabilities by incorporating rules

directly into databases to support inferences. To develop FOO, I employed the Linked Open Terms (LOT) methodology [211], chosen for its alignment with our project's needs, including the ability to model natural language statements and support the publishing and maintenance of the ontology. The development of the ontology progressed through iterative stages, including requirement gathering, implementation, evaluation, and publication. Figure 3.1 illustrates the comprehensive lifecycle of FOO's development process. FOO's actors play distinct roles during its life cycle. Bio-scientists contribute expertise on Sabah's forest wildlife, including Bornean elephants and their endangered status. Wildlife researchers provide real-world scenarios, such as elephants' movements and incidents. Ontology developers focus on integrating FooDS with AI-enabled frontends using SPARQL queries. Computer scientists highlight the benefits of Linked Data stores (i.e., FooDS) in bridging data silos. Data scientists assess FooDS usability and explore applications of machine learning and deep learning.

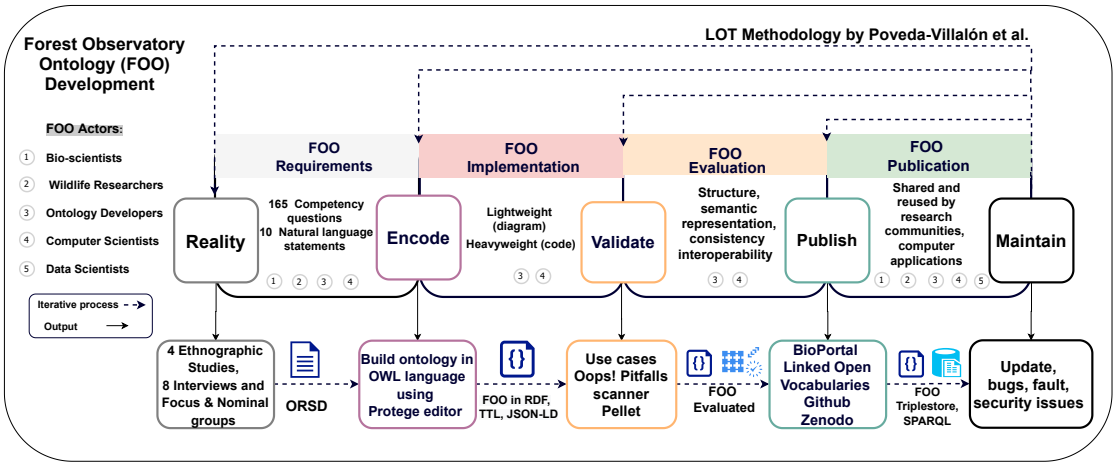


Figure 3.1 FOO Ontology Development phases, inspired by Linked Open Terms (LOT) methodology [211].

3.2.1 Ontology Requirements

During the initial phase of the ontology development process, the Ontology Requirements Specification Document (ORSD) [240] was crafted, adhering to the guidelines outlined in the LOT methodology. The ORSD outlines critical details, such as the ontology's scope, intended purpose, and the use cases it aims to support. This phase actively involves domain experts identifying use cases for the ontology and selecting the datasets that will be modelled. Competency Questions (CQs) in Table .3 and 3.3, Natural Language Statements (NLSs) in Table 1 were compiled and various use cases were created for biologists and wildlife

researchers. CQs, as defined in [112], outline the functional requirements of the ontology by formulating questions that the ontology should answer using query languages. NLSs are short affirmative phrases that convey information to be included in the ontology. Use cases describe real-world scenarios that the proposed ontology aims to address.

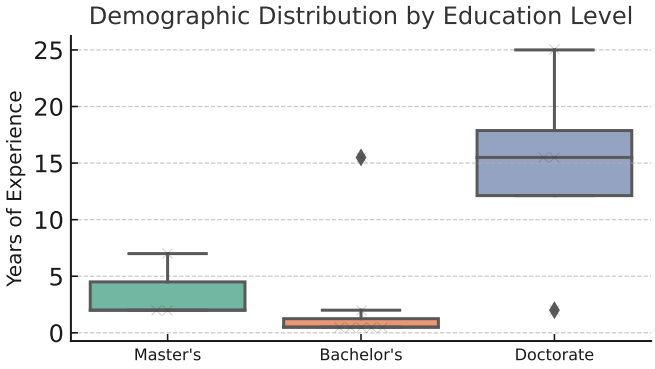
To meet these requirements, I engaged in three distinct activities: The first activity was an *ethnography* to gain insight into the wildlife research community, informed by casual interviews and observations during data collection. The second involved conducting semi-structured *interviews* with eight wildlife researchers from Cardiff University in Wales and the DGFC in Sabah, Malaysian Borneo. I organised a text-based focus group for the third activity and conducted a nominal group technique session. Three administrative documents were created for each activity: participants' information sheets, consent forms, and demographic questionnaires. Participants' information sheets briefly outlined the project objectives and procedures for the activities. Consent forms enabled us to obtain signed permission from the participants to proceed with the activities. Finally, the demographic questionnaire collected non-personal details from participants, such as their education level, occupation, and years of experience.

Ethical clearance (COMSC/Ethics/2021/039) for collecting the necessary data was granted by Cardiff university's research ethics board. The participants were provided with both online and paper versions of the documents. To enlist participants, snowball sampling technique was followed [91] and collaborated with DGFC. The first meeting was with biologists within Cardiff University network, requesting them to refer to individuals who aligned with the study criteria. Although the target was six participants for formal interviews, this direct engagement strategy proved effective, and semi-structured interviews with eight participants were successfully conducted. In the discussion groups (nominal and focus), at least five participants joined each group, receiving responses from 14 participants via Google Forms in total. Interviews were transcribed using Microsoft Word, written notes were taken for the discussion groups, and the ethnographic studies were summarised. In line with methodologies adopted in prior research [78], the data gleaned from interview transcriptions and discussion group notes were manually coded into Competency Questions (CQs) (see Table .3) to guide the ontology's development.

Figure 3.2 shows the demographic details of the study participants.

Ethnography

Ethnographic research at DGFC were exercised in the summer of 2022 to observe the collection and processing of wildlife data [163] (Figure 3.3). Four activities were carried out: (i) comparing butterfly diversity, (ii) comparing Proboscis Monkey Activity, and (iii) finding



	count	mean	std	min	25%	50%	75%	max
Bachelor's	7.0	2.86	5.6	0.5	0.5	0.5	1.25	15.5
Doctorate	4.0	14.5	9.46	2.0	12.12	15.5	17.88	25.0
Master's	3.0	3.67	2.89	2.0	2.0	2.0	4.5	7.0

Figure 3.2 Participants' Demographic Information

the tracked Sunda pangolin. (iv) finding the Elephas maximum (Asian elephants). These ethnographic experiments were conducted to understand cultural and operational dynamics, providing crucial insights to guide research and the design of the proposed ontology. For example, they revealed how data are collected and processed, identified the species present in the forests of Sabah, and highlighted some use cases that can be tailored to real-world needs.

1. Comparing butterfly diversity: The ethnographic research contrasted butterfly populations in a tropical rainforest and an oil palm plantation. The rainforest revealed rich biodiversity with 372 butterflies across 23 species, notably G. harina, constituting 67% of its butterfly fauna. In contrast, the plantation had only nine butterflies of six species, with no unique species. The findings, showing a mean species richness of 8.4 in the rainforest and 2.2 in the plantation, highlight the significant impact of habitat on butterfly diversity.

2. Comparing Proboscis Monkey Activity: The ethnographic study conducted by BSc students from Bioscience, with assistance from staff wildlife researchers at the Danau Girang Field Centre (DGFC). This study focused on variations in the behaviour of proboscis monkeys across different time periods (multiple days). The study used two methodologies: visual surveys along the Kinabatangan River and nearby forest trails and bio-acoustic monitoring facilitated by AudioMoth devices. Preliminary visual data suggested a higher sighting frequency of monkeys along the riverbank than in the forest, somewhat challenging the initial hypothesis of peak afternoon activity at the river. However, the acoustic data, encompassing over 22 hours of recordings from

each device, are yet to be analysed and will be crucial in validating or refuting the hypothesis regarding the unpredictability of activity patterns. Despite its insights, the study acknowledged several limitations, including adverse weather conditions, human observation constraints, and the possibility of repeated sightings of the same monkeys.

3. Finding the tracked Sunda pangolin: In the early hours, I entered the forest with fellow researchers, equipped for protection against insects and rain, aiming to locate the Sunda pangolin using a noise-emitting antenna designed for proximity detection. This method is critical for tracking species in extensive forested areas. The increasing strength of the antenna's signal indicated our approach to the pangolin, a species known for its effective camouflage and quiet movement.

4. Finding the Asian elephants (Elephas Maximum): Leaving Sandakan Jetty, me and my colleagues headed towards the DGFC centre through the dense forest. On our boat ride, we spotted a group of Asian elephants swimming in the lower Kinabatangan River.

Figure 3.3 Forest Observatory Ontology Development's activities collage

Interviews

Eight semi-structured face-to-face discussions with specialists in genetics and biology were conducted, focusing on wildlife conservation. Seven interviews were based in Sabah (Malaysian Borneo), apart from one individual from the United Kingdom who volunteered in the DGFC. The participants had diverse experience in landscape ecology and conservation biology research, with their expertise spanning from one to twenty-five years. To recruit participants for the study, the first bioscientist was identified from our university mailing list, and then snowballing technique was used to identify and recruit participants from hard-to-reach populations [181]. The process begins by identifying a participant through an internal mailing list, for instance. This participants then refers others from their networks who also fit the criteria. These nominees were determined with an information sheet about the interview and details of ethics approval two weeks before their interviews. During these sessions, a

consistent semi-structured guide was followed to delve into the types of data the participants collected and processed and their aims to use them to make well-informed decisions swiftly. Every participant completed their interview within 60 minutes, and the audio recordings were preserved for detailed analysis. The interview questions covered a range of topics, including:

- What is your opinion about a given User Interfaces mock-up?
- What features would you like to use?
- What is your feedback about the delivered linked data store prototype/ outcome?
- What are the types of collected data?
- How do you process the collected data?
- What are the tools and methods used to process the data?
- How do you access and interact with the data?
- What are the drawbacks of your current data system?
- What questions do you require your data environment to answer?
- What would the ideal data model look like for you (e.g., chronological data catalogue, interactive interface with links to downloadable datasets)?

Given that DGFC faces data silos due to project-specific data collection practices. These interview questions aim to identify strategies for breaking these silos, integrating datasets, enhancing data discovery, and providing organised, accessible data to support informed decision-making.

Interviews analysis and findings

Analysis of the interview transcriptions was conducted using inductive coding [27]. This approach entails thoroughly examining the data, including interview transcripts, field notes, and documents, to identify text segments of interest or significance. Each segment was labelled in a manner that mirrored the participants' own words or specific details of the data. As the analysis progressed, these initial labels were aggregated into broader themes that naturally arose from the dataset. The analysis was cyclical; I kept comparing new data to existing codes and themes and refining them as needed. Table 3.1 presents the themes that emerged from this process, along with their descriptions. The research findings revealed

a collective desire for improved data management, visualisation, and accessibility across wildlife research activities. Studies ranging from animal tracking to vegetation studies have highlighted the demand for simple, unified, and user-friendly interfaces for data management. Interviewees expressed challenges with manual data entry, integrating disparate data sources, and the need for better tools to visualise and analyse data, mainly through maps for spatial understanding. Key quotes reflecting these themes include:

- Participant(2): "*All of this raw data I keep it myself like I save it in my external hard drive as well*" indicating challenges with data accessibility and sharing.
- Participant(5): "*We don't have it in the GPS, in the camera traps, but since I was advised us to do so, we have now labelled the pictures in the timestamp of the name and using the name.*" showing efforts to improve data organisation but still highlighting manual processes.
- Participant(7): "*So you might want to say into Google, like where are the elephants right now or where have the elephants been in the last two weeks?*" This illustrates the need for intuitive data query methods that can provide real-time or specific historical insights based on natural language processing.

These findings highlight the need for wildlife research platforms that integrate diverse data sources, improve contextual data access, and enable efficient, complex query resolution. The ideal system should manage and visualise current data, such as GPS tracking, whilst adapting to evolving conservation needs by incorporating new data types and analytical methods.

Focus and nominal groups

In the form of visual materials, a map of Sabah, Malaysian Borneo was created, displaying diverse types of wildlife data, such as elephant movements. Three information cards were printed detailing the GPS collar, soil sensor, and vegetation data, with blank spaces for note-taking and participant comments. Over two consecutive days, nominal and focus groups were held, with six members and one moderator in the former and seven members in the latter. During the nominal group, participants brainstormed ideas in response to the study questions, wrote them down, and then shared them with the group in a round-robin manner. For the focus group, participants received a copy of the primary map and three data-type cards. They were requested to suggest ideas, potential use cases, and questions that could be addressed using these different datasets. From the discussion groups, a list of use cases was collected for FOO and the relevant datasets, exploring their potential applications and usefulness in

Table 3.1 Overview of Participant Feedback Themes. This table outlines the main feedback themes from evaluating our proposed data management system, including design impressions, functional requirements, and user experiences.

Theme	Description
Design	Participants' impressions of the mock-up's ease of use, visual appeal, and overall user experience.
Functional Requirements	Features participants find essential or desirable for their work, such as data visualisation tools, search functionality, or customisation options.
Data Diversity	The variety and nature of data that participants deal with, including qualitative, quantitative, temporal, or spatial data.
Analytical Methods	How participants process data, including data cleaning, analysis techniques, and transforming raw data into usable information.
Technology	Software, tools, and methods used for data processing, highlighting preferences, effectiveness, and limitations.
Usability	How participants access, explore and manipulate data, including databases, APIs, or interactive dashboards.
Challenges	identified issues with current data systems include lack of integration, poor usability, or inadequate functionality.
Prototype Evaluation	Participants' assessments of the prototype's functionality, performance, and how well it meets their needs or expectations.
Desired Outcomes	The Specific questions or problems participants need their data environment to address, reflecting on gaps in current systems.
Vision for the Future	Participants' conceptualisation of the ideal data model or system.

informed decision-making. This exercise has been a significant portion of the Competency Question (CQs) and the Natural Language Statements (NLSs) for ontology, offering valuable insights into its development and application. Participants gathered various perspectives and ideas, including those new to such activities, resulting in a rich collection of spoken and written information. Both sessions were conducted ethically, with consent obtained, and video recordings were recorded. Subsequently, the Ontology Requirements Specification Document (ORSD) in Appendix .1, the ontology development sheet containing Competency Questions (CQs) listed in Tables .3 and 3.3, and the Natural Language Statements (NLSs) in Table 1 were finalised, with the corresponding SPARQL queries added in Appendix .3. The methodology details are uploaded to the ontology website (ontology.forest-observatory.cardiff.ac.uk).

3.2.2 Ontology Implementation

Based on the Ontology Requirements Specification Document (ORSD) crafted in the requirements phase, It was deduced that FOO's scope would include Internet of Things (IoT) concepts, such as sensors and their observation, and wildlife aspects, like animals. For example, the datasets of interest and proposed use cases include data from *"sensors"* monitoring *"animals"* and *"land"*. For instance, an animal GPS collar tracks an elephant, recording

	Open Data		Sensor Data	
Competency Questions (CQs) SPARQL in Appendix I .3	Soil	Veg	GPS	Answer
CQ1 Where do elephants forage?		*	*	✓
CQ2 What are the daily movement patterns for Elephant X in June??			*	✓
CQ3 What are the yearly movement patterns for Elephant X??			*	✓
CQ4 How do the movements of Elephant X relate to human and urban areas?			*	✓
CQ5 Has elephant x died?			*	✓
CQ6 Why has elephant x died?				✓
CQ7 What are the suitable environmental conditions for elephant x to survive?	*	*	*	x
CQ8 What can we learn from the movements of Elephants X, Y, and Z?			*	✓
CQ9 How does Elephant X use Habitat Site Y??		*	*	✓
CQ10 What is the range of habitat sites used by Elephants X, Y, and Z?		*	*	✓
CQ11 Where was Elephant X located during the flood season in the Lower Kinabatangan area??			*	✓
CQ12 What was the average speed of Elephant X during the flood season?			*	✓
CQ13 Is Elephant Dara near (5 Km) the danger zone (poachers' area) today?			*	✓
CQ14 How did Elephant X's movements change with climate change in 2014?			*	✓
CQ15 What are Elephant X's preferred habitats based on prolonged stays in areas?	*		*	✓
CQ16 How far was Elephant X from the oil plantation fencing?			*	✓
CQ17 When was Elephant X near the oil plantation fencing?			*	✓
CQ18 What is the distance traveled between each of Elephant X's stops (sleeping)?			*	✓
CQ19 * Which elephants met this month?			*	✓
CQ20 Which sites were revisited by Elephant X month?			*	✓
CQ21 What environment or habitat does Elephant X prefer, based on the prolonged time spent in a certain area?	*	*	*	✓
CQ22 Was there any significant change in Elephant X's movement patterns between June and July 2012?			*	✓
CQ23 Has Elephant X visited Village Y in year Z?			*	✓
CQ24 What is the movement range of Elephant X during Month Y?			*	✓
CQ25 What is Elephant's activity (speed) during Month Y?			*	✓
CQ26 Are there any interactions between collared elephants during the flood season?	*		*	✓
CQ27 What is the status of Elephant X's tracking collar battery?			*	✓
CQ28 What habitat has Elephant X selected this season?	*	*	*	✓
CQ29 What is the average elevation of Elephant X during a specific time range?			*	✓
CQ30 Which elephant came near the logged site?		*	*	✓
CQ31 Which elephant came near the semi-logged site?		*	*	✓
CQ32 Which elephants crossed the river?			*	✓
CQ33 What is the canopy height for the distance traveled by Elephant X during the flood season?	*	*	*	✓
CQ34 Which elephants are near the oil palm plantations this week?			*	✓
CQ35 What is the home range for all collared elephants?			*	✓
CQ36 What is the distance traveled by Elephant Y over a specific period?			*	✓
CQ37 What are the altitudes of the collared elephants?			*	✓
CQ38 What are the body/environment temperatures for collared elephants?			*	✓
CQ39 What is the behavior of Elephants X and Y this month?			*	✓
CQ40 Does Elephant X need help?			*	✓
CQ41 What are the distribution patterns of Elephants X and Y during this month?			*	✓
CQ42 Are Elephants X and Y's favorite foods in a particular area?		*	*	x
CQ43 Do we need to create corridors along rivers/palm plantations, or is it not an obstacle for elephants to cross the river?			*	x
CQ44 Why have the elephants' collars been fitted for almost two years?			*	✓
CQ45 What are the migration patterns of Elephants X during the flood season?			*	✓
CQ46 What are the favorite locations that Elephant X likes to visit during certain times of the year?			*	✓
CQ47 Where are elephants likely to come into contact with humans?			*	✓
CQ48 What are the places where elephants may be vulnerable?			*	✓
CQ49 Where can we assign locations to rangers?			*	✓
CQ50 How to track (investigate) the last location of a dead elephant?			*	✓

Table 3.2 Competency Questions (CQs) extracted from research activities such as ethnographic research, interviews, and nominal and focus groups

	Open Data			Sensor Data
Competency Questions (CQs)	*Soil*	*Veg*	*GPS*	*Answer*
CQ51 * Will the elephants be arriving at DGFC soon?			*	✓
CQ52 How many satellites did the collar detect? (COV=0, speed=0)			*	✓
CQ53 Which elephants are close to the river today?			*	x
CQ54 Which elephants are close to oil plantations?			*	✓
CQ55 Which elephant roams near the Sabahmas site?			*	✓
CQ56 Which elephant roams near small steep sites?			*	✓
CQ57 Which elephant is likely to visit Ribubonus, kg. Kiabau, and Reka Halus 12ha?			*	✓
CQ58 What locations could have snares?			*	x
CQ59 * Is Elephant X sick, injured, or dead?			*	✓
CQ60 Which elephants are likely to conflict with humans?			*	x
CQ61 What is the soil condition during certain times of the year?	*			✓
CQ62 What types of soil are available throughout the year? Dry, muddy, swamps.	*			✓
CQ63 What are the locations (soil type) that elephants prefer?	*			x
CQ64 What are the mineral content (salt and others) in a particular location?	*			✓
CQ65 Is there any metal in the soil in that area?	*			x
CQ66 What are the chemical and agro-chemical concentrations in the soil of a certain area?	*			✓
CQ67 Does the soil in location x contain disease pathogens?	*			✓
CQ68 Which area needs pesticide spraying?	*			✓
CQ69 What is the soil moisture level in a specific location?	*			✓
CQ70 What is the presence of minerals in the soil?	*			✓
CQ71 Are there signs of heavy metals in the soil?	*			x
CQ72 * Where are the salt licks located?	*			✓
CQ73 What are the mineral and salt concentrations in the soil that indicate the presence of salt licks in a particular location?	*			✓
CQ74 What is the pH level of the soil?	*			✓
CQ75 What is the temperature reading from the soil sensor?	*			x
CQ76 What is the soil moisture in a certain location?	*			x
CQ77 Is the soil in this area healthy for animals?	*			✓
CQ78 Is the soil fertile in this area?	*			✓
CQ79 What is the moisture rate of the soil in this area (i.e., provide geo-location)?	*			x
CQ80 Where to plant crops for elephants (i.e., soil moisture rates)?	*			✓
CQ81 Could planting in safer areas (healthy soil) influence animal movements?	*			✓
CQ82 Could we predict crop yield based on soil data?	*			x
CQ83 What soil metrics help us predict flooding?	*			✓
CQ84 What are the metrics of healthy soil with less/no chemical pollution from oil palm plantations?	*			✓
CQ85 Why do elephants not like to walk on wet soil (movement prediction)?	*			✓
CQ86 What are the chemical levels of the soil in Protected Area 1?	*			✓
CQ87 What are the soil nutrient levels?	*			✓
CQ88 What is the effect of moisture on nutrients and oxygen levels?	*			✓
CQ89 What is the ideal soil moisture rate for an elephant to give birth?	*			✓
CQ90 What are the soil conditions in the areas that have elephant grass?	*			✓
CQ91 How to conserve suitable soils for the elephants to have food in the future?	*			✓
CQ92 What soil moisture do elephants spend most time on?	*			✓
CQ93 What do elephants eat?		*		x
CQ94 Where do bamboo shoots grow?		*		x
CQ95 Where could we find areas with the inner trunk of oil palms?		*		✓
CQ96 Where could we find areas with broad leaves?		*		✓
CQ97 Where could we find areas with vines?		*		x
CQ98 How can vegetation and site habitat information help understand the future patterns/locations of elephants?		*		x
CQ99 Do elephants drink lots of water?			*	x
CQ100 Where do we find fruit farms in lower Kinabatangan?		*		x
CQ101 What areas have fewer trees?		*		✓
CQ102 What plant species to conserve in the areas the elephants visit?		*	*	x
CQ103 What plant species effected by deforestation?		*		x
CQ104 Which plant species are cultivated by the Grow Borneo project?		*		✓
CQ105 How many trees has the Grow Borneo project planted in the last five years?		*		x

Table 3.3 Competency Questions (CQs)-Continued

	Open Data		Sensor Data	
Natural Language Statements (NLSs)	*Soil*	*Veg*	*GPS*	*Answer*
NLS1 Tracking elephant locations so that the wildlife department can give warnings to local people about the arrival of elephants.			*	x
NLS2 Examples of areas with elephant grass (Nappier), other grasses, bark, palm shoots, young leaf trunks, soft plants, and bananas.		*		x
NLS3 Focus on the area of Lower Kinabatangan and the 14 collared elephants living there.			*	✓
NLS 4 Collared elephants will not go to primary forest sites.		*	*	x
NLS5 The datasets in this research could be used to generate predictions.			*	x
NLS6 Elephants do not intend to cause damage. It may occur when their strong and huge bodies come in contact with things.			*	x
NLS 7 Nearly all wild pigs in the area of Kinabatangan died from influenza viruses.				x
NLS 8 There was a famous story about the rhino who lost one leg from poaching. It survived on three legs for a long time.				x
NLS9 Female Asian elephants are tusk-less.			*	x
NLS10 Male Asian elephants are more likely to explore human areas than females, attracted by food.			*	CQ30

Table 3.4 Natural language statements and what data set can fulfil the task.

geographic location observations at different and equally spaced time intervals and temperature readings at each specified interval. Searching in scholarly resources and ontology repositories, relevant ontologies were identified. The search included Google Scholar [105], BioPortal repository, and other pertinent websites. The selection criteria stipulated that publications must be published between 2015 and 2020. Variety of search terms were used, such as "sensor data ontology," "semantic modelling for sensor data," "semantic IoT data," and "IoT ontology." Several domain-specific ontologies were found for modelling sensors and wildlife data. The Semantic Sensor Network (SSN) ontology describes the sensory observation processes (SSN, SSN2). Within SSN Version 2, a Sensor, Observation, Sample, and Actuation (SOSA) ontology is suitable for lighter use without the whole SSN [132]. IoT-Lite ontology provides foundational descriptions of IoT resources, whilst the Smart Applications REFerence (SAREF) ontology focuses on referencing IoT appliances [61]. The Extensible Observation Ontology (OBOE) [165] models terms, such as `observation` and its `measurement`. For wildlife ontologies, notable examples include GeoSpecies ontology, BBC Wildlife Ontology (BBC-WO) (bbc.co.uk/ontologies/wildlife-ontology/), the African wildlife ontology [140], and an ontology of core ecological entities named Ecocore, catering to specific wildlife aspects depending on the intended purpose and use case. As a result, most commonly used ontologies for modelling sensor data were filtered out, such as the SSN [256]. Among the shortlisted ontologies is the SAREF ontology [61], which is designed for smart appliances, IoT devices, and services; however, it may not adequately model sensor data observations. IoT-Lite ontology [24] provides a basic framework of classes and properties for describing IoT devices, sensors, and actuators. However, for our specific use cases, more

classes are needed to model the sensor's observations and associated properties than just the sensors.

The W3C Web of Things (WoT) ontology (w3.org/TR/wot-thing-description11/) is a flexible and modular ontology that can be customised to fit different use cases, allowing for interoperability across various IoT systems and domains. Although it covers distinct aspects of IoT devices and services, its flexibility and generality can make adapting to our specific requirements challenging. For instance, the 'Thing' class in WoT models the IoT device, service, or data source, whereas the 'Sensor' class is better suited for modelling sensor observations. The FIESTA-IoT ontology (iot.ee.surrey.ac.uk/ontology/fiesta-iot.owl) primarily models IoT-related concepts but includes more entities than needed. It incorporates classes from the SSN ontology (Version 1) [52], W3C Web of Things (WoT) Thing Description, and oneM2M standard (onem2m.org). The IoT-Semantics Ontology is another flexible ontology; however, its lack of sufficient documentation makes it challenging for developers to adapt.

Ontology Resue

After conducting an in-depth examination and comparison of contemporary ontologies, concepts ere reused from SSN ontology (Version 2) [256]. This ontology distinguishes itself from its modular structure, comprising three integrated ontologies: the original SSN ontology (Version 1), the Sensor, Observation, Sample, and Actuator (SOSA) ontology [132], and the Quantities, Units, Dimensions, and Types (QUDT) ontology.eth Such integration makes the SSN ontology (Version 2) well-suited to our needs. FOO extracted SOSA ontology from SSN version 2 using owl:imports, reused classes with 'owl:equivalentClass', and properties with 'owl:equivalentProperty'. FOO directly uses the geolocation points, specifically longitude and latitude, from the W3C's Basic Geo (WGS84 lat/long) Vocabulary available at (w3.org/2003/01/geo/).

FOO extends the BBC Wildlife Ontology (WO) by reusing its taxonomic structure through equivalence relations and defined hierarchies. At the kingdom level, $foo : Animalia$ is specified as $foo : Animalia \equiv wo : Animalia$ and $foo : Animalia \sqsubseteq$ owl:Thing, representing all animals. Moving to the phylum level, $foo : Chordata$ is defined as $foo : Chordata \equiv wo : Chordata$ and $foo : Chordata \sqsubseteq foo : Animalia$, encompassing vertebrates and related taxa. The class level includes $foo : Mammalia$ ($foo : Mammalia \equiv wo : Mammalia$, $foo : Mammalia \sqsubseteq foo : Chordata$) for mammals and $foo : Reptilia$ ($foo : Reptilia \equiv wo : Reptilia$, $foo : Reptilia \sqsubseteq foo : Chordata$) for reptiles. Within $foo : Mammalia$, orders such as $foo : Proboscidea$ ($foo : Proboscidea \equiv wo : Proboscidea$, $foo : Proboscidea \sqsubseteq foo : Mammalia$) for elephants and $foo : Carnivora$ ($foo : Carnivora \equiv wo : Carnivora$, $foo : Carnivora \sqsubseteq foo : Mammalia$) for carnivorous mammals are included. Similarly, un-

der $foo : Reptilia$, the order $foo : Squamata$ is defined as $foo : Squamata \equiv wo : Squamata$ and $foo : Squamata \sqsubseteq foo : Reptilia$, representing snakes and lizards. At the family level, FOO introduces $foo : Elephantidae$ ($foo : Elephantidae \equiv wo : Elephantidae$, $foo : Elephantidae \sqsubseteq foo : Proboscidea$), encompassing elephants and related extinct species. Additionally, general taxonomic categories such as $foo : Genus$ ($foo : Genus \equiv wo : Genus$) and $foo : Species$ ($foo : Species \equiv wo : Species$) enable further classification. By reusing and aligning with WO, FOO ensures semantic coherence ($\equiv$), a well-structured hierarchy ($\sqsubseteq$), and adaptability for domain-specific needs.

FOO adopts and extends key classes and properties from SOSA to model observations, sensors, and their relationships in a semantically rich framework. The $foo : Observation$ class is defined as equivalent to $sosa : Observation$ ($foo : Observation \equiv sosa : Observation$) and represents the act of estimating or calculating the value of a property of a $foo : FeatureOfInterest$, such as an elephant or tree. Similarly, $foo : Sensor$ ($foo : Sensor \equiv sosa : Sensor$) describes devices, agents, or software involved in implementing a procedure. Observable qualities are represented by $foo : ObservableProperty$ ($foo : ObservableProperty \equiv sosa : ObservableProperty$), which denotes measurable characteristics such as temperature or speed, while $foo : FeatureOfInterest$ ($foo : FeatureOfInterest \equiv sosa : FeatureOfInterest$) represents the entity being observed. FOO also reuses and aligns several SOSA object properties. The property $foo : hasFeatureOfInterest$ ($foo : hasFeatureOfInterest \equiv sosa : hasFeatureOfInterest$) links an observation to its feature of interest, while its inverse, $foo : isFeatureOfInterestOf$ ($foo : isFeatureOfInterestOf \equiv sosa : isFeatureOfInterestOf$), connects the feature to its corresponding observations. The property $foo : madeBySensor$ ($foo : madeBySensor \equiv sosa : madeBySensor$) relates an observation to the sensor that produced it, with the inverse relation defined as $foo : madeObservation$ ($foo : madeObservation \equiv sosa : madeObservation$). Furthermore, $foo : observedProperty$ ($foo : observedProperty \equiv sosa : observedProperty$) connects observations to the specific properties being measured, and its inverse, $foo : isObservedBy$ ($foo : isObservedBy \equiv sosa : isObservedBy$), relates observable properties to the sensors capable of detecting them. The $foo : observes$ property ($foo : observes \equiv sosa : observes$) establishes a direct relationship between a sensor and the observable property it monitors. These alignments ensure semantic interoperability with SOSA while enabling FOO to extend and specialise these concepts for wildlife observation and environmental monitoring.

FOO introduces a range of domain-specific classes to represent wildlife species, each structured within a well-defined taxonomic hierarchy and aligned with external references using description logic. The class $foo : Primates$ is defined as $foo : Primates \sqsubseteq foo : Mammalia$, representing an order of mammals characterised by large brains and including

species such as lemurs, monkeys, and hominids. Within this, $foo : Cercopithecidae$ is defined as $foo : Cercopithecidae \sqsubseteq foo : Primates$ to represent African and Asian monkeys, further narrowed to the genus level with $foo : Nasalis$ ($foo : Nasalis \sqsubseteq foo : Cercopithecidae$), which includes the proboscis monkey. The species $foo : NasalisLarvatus$ is defined as $foo : NasalisLarvatus \sqsubseteq foo : Nasalis$ and is aligned with external taxonomy via $owl : equivalentClass$

$foo : ElephasMaximus$ (Asian elephant) is represented as $foo : ElephasMaximus \sqsubseteq foo : Elephantidae \sqcap foo : FeatureOfInterest$, integrating its ecological and conservation significance. For reptiles, $foo : Pythonidae$ is defined as $foo : Pythonidae \sqsubseteq foo : Squamata$ and aligned with external taxonomy via $owl : equivalentClass$. This lineage includes $foo : Malayopython$ ($foo : Malayopython \sqsubseteq foo : Pythonidae$) and its subclass, the reticulated python $foo : MalayopythonReticulatus$ ($foo : MalayopythonReticulatus \sqsubseteq foo : Malayopython \sqcap foo : FeatureOfInterest$), which is further aligned with external references. The class $foo : ManisJavanica$ (Sunda pangolin) is defined as $foo : ManisJavanica \sqsubseteq foo : Mammalia \sqcap foo : FeatureOfInterest$ and aligned with $owl : equivalentClass \langle http : //purl.bioontology.org/ontology/NCBITAXON/9974 \rangle$. This critically endangered species is described with details on its unique keratin armour and its role in controlling insect populations. These classes, enriched with external references and conservation-focused descriptions, exemplify FOO's capability to integrate domain-specific wildlife data within a semantically rich framework.

FOO defines a set of data properties to describe observational, geographical, and environmental measurements, ensuring semantic clarity and interoperability. Data properties for GPS observations include $foo : temperature$ ($foo : temperature \sqsubseteq foo : gPSObservation \sqcap xsd : double$), which captures the temperature in Celsius during data collection, and $foo : direction$ ($foo : direction \sqsubseteq foo : gPSObservation \sqcap xsd : integer$), representing the directional movement of observed features. Spatial data properties such as $foo : latitude$ and $foo : longitude$ link to $pos : lat$ and $pos : long$ respectively, ensuring consistency with geospatial ontologies. Properties like $foo : gMTDate$ and $foo : gMTTime$ define temporal metadata for data logging, aligned to global time standards ($foo : gMTDate \sqsubseteq xsd : date$). For soil observations, properties such as $foo : clay$, $foo : silt$, and $foo : soilPH$ ($foo : soilPH \sqsubseteq foo : soilObservation \sqcap xsd : double$) capture detailed soil composition and quality data, essential for ecological studies. Tree observations are supported by properties like $foo : treeDBH\_cm$ and $foo : treeHeight\_m$, representing diameter at breast height and tree height ($foo : treeHeight\_m \sqsubseteq foo : treeObservation \sqcap xsd : float$). Camera trap image data properties include $foo : imageFile$ and $foo : path$ ($foo : path \sqsubseteq foo : imageObservation \sqcap xsd : anyURI$), defining the storage location and metadata of captured wildlife images. The

foo : *animalDetected* property links image observations to identified species, while foo : *cameraLocation* contextualises image data geographically. These properties integrate diverse observational and environmental metrics, enabling FOO to provide a detailed, interoperable framework for wildlife and environmental monitoring.

FOO also defines a detailed set of instances for sensors and observed features to contextualise wildlife monitoring data. Sensor instances such as foo : *aqeelaGPS*, foo : *bikang*1*GPS*, and foo : *bikang*2*GPS* are represented as *owl* : *NamedIndividual* $\sqcap$ foo : *Sensor* and are explicitly associated with specific elephants (foo : *hasFeatureOfInterest* links them to foo : *Aqeela*, foo : *Bikang*1, and foo : *Bikang*2, respectively). These sensors also observe a range of properties through their association with foo : *gPSObservation*, enabling the collection of real-time spatial and behavioural data. Individual elephants are instantiated as *owl* : *NamedIndividual* $\sqcap$ foo : *ElephasMaximus*, where each, such as foo : *Aqeela* or foo : *Sejati*, is defined with unique labels and *skos* : *definitions* to capture contextual information. For example, foo : *Aqeela* is defined as a female Asian elephant, while foo : *Guli* represents a male counterpart. This structure allows for detailed semantic representation and tracking of individual animals in ecological studies. Additional feature instances include foo : *Soil* (*owl* : *NamedIndividual* $\sqcap$ foo : *FeatureOfInterest*), observed through foo : *soilObservation*, which captures soil-specific properties like foo : *soilPH* and foo : *totalC*. Similarly, foo : *Tree* represents arboreal features, with properties like foo : *treeHeight_m* and foo : *treeDBH_cm* observed through foo : *treeObservation*. Camera trap images, represented by foo : *Image*, are associated with foo : *imageObservation* and link to specific metadata like foo : *imageFile* and foo : *cameraLocation*. By employing such well-defined instances and linking them to their respective observations and features of interest, FOO facilitates granular, interoperable representation of ecological monitoring data. This approach not only enhances the semantic clarity of the ontology but also ensures its utility in wildlife conservation efforts.

FOO (see Appendix .4) contains 81 classes, 73 properties, and 176 individuals. Table 3.5 lists a small part of its content, including concepts that represent wildlife data generated by sensors and extracted from data collected during the ontology requirement phase. Specifically, FOO includes data on wildlife species and devices observed during ethnography, such as the Asian elephant, Sunda Panogolin and Proboscis Monkey. Figure 3.4 illustrates part of FOO data modeling. FOO models ‘foo:jasmin‘ as an instance of the class ‘foo:ElephasMaximus‘ (representing the Asian elephant), which is a subclass of ‘foo:Elephantidae‘, further subclassed under ‘foo:Mammalia‘, and ‘foo:Chordata‘. ‘foo:jasmin‘ is identified as the ’Feature of Interest’ for ‘foo:gPsObservation‘, an instance of ‘foo:Observation‘, which is made by ‘foo:Sensor‘. The light pink classes indicate elements imported and mapped from SOSA,

while the light yellow classes represent those imported and reused from the BBC Wildlife Ontology (WO). The light blue spatial box are the directly-used basic Geo vocabulary. Further, Figure3.5 illustrates a collapsed heavy weight version of FOO, visualised using WebVowl (service.tib.eu/webvowl/).

3.2.3 Ontology Evaluation

Various ontology evaluation techniques were investigated and discovered that ontology evaluation primarily focused on assessing quality and accuracy during and after its development [203]. Raad et al. [213] identified four ontology assessment methods from the literature: (i) gold standards, (ii) corpus-based, (iii) criteria-based, and (iv) task-based. McDaniel et al. [172] described ontology evaluation as a two-fold process, namely, the glass-box and black-box approaches. The former evaluates the ontology incrementally throughout its life-cycle, also known as component evaluation. By contrast, the latter is a task-based approach to evaluating an ontology's performance in a specific task or application [173]. The most suitable method for evaluating an ontology depends on the intended purpose. The proposed ontology, FOO, is designed to support applications integrating heterogeneous data sources for decision-making. Thus, the *structure, semantic representation, and interoperability* were evaluated. To assess the structure and semantic representation, open-source online scanner Foops! (foops.linkeddata.es) was selected, as the baseline for evaluating semantic representation, Pellet [231] to detect any inconsistencies via reasoning and, SPARQL queries to evaluate FOO's accuracy and efficiency. Subsequently, in the following section, the black-box (i.e., task-based) approach was followed to assess the applicability and interoperability of knowledge graphs (FOODS), focusing on how well it addresses use cases and their efficiency in data exchange between different computer systems.

FOOPS! Evaluation

FOOPS! [95]is a web service created to evaluate the compliance of vocabularies and ontologies with the FAIR principles, making sure that the ontology under evaluation is *findable, accessible, interoperable and reusable*. FOOPS! performs a series of checks to ensure compliance with the FAIR principles. Regarding the *Findable* dimension, it does nine checks to see if the ontology URI is persistent and resolvable, has a version IRI that is unique for each version and has at least the title and description of the instance. Moreover, it verifies if the ontology prefix and namespace are registered in external registries like prefix.cc and LOV. In the *Accessible* dimension, three checks ensure proper content negotiation with at least one RDF serialisation and HTML format and verify that the URI protocol is open

Table 3.5 Main Forest Observatory Ontology, classes and descriptions

Class/Properties	Label	Description < foo:http://w3id.org/def/foo#>
foo:ElephasMaximus	Asian Elephant	Elephas maximus, commonly known as the Asian elephant, is a species of large mammal native to various regions in South and Southeast Asia, including India, Sri Lanka, Thailand, and parts of Indonesia. It is distinguished by its smaller ears compared to its African relatives, and it has a prominent domed head with two hemispherical bulges. The Asian elephant is classified as Endangered due to significant threats from habitat loss, fragmentation, and poaching. This species plays a crucial ecological role, aiding in forest maintenance through seed dispersal and the creation of clearings in dense vegetation. Bornean elephants exhibit distinct morphological and behavioural traits compared to mainland Asian elephants, and their genetic uniqueness emphasises their priority for conservation efforts. Although they are considered an evolutionary significant unit requiring tailored conservation measures, their formal recognition as a subspecies awaits further research. Restricted to about 5% of Borneo, primarily in Sabah, Bornean elephants typically form family groups of 5 to 20 individuals, occasionally merging into larger herds of up to 200.
foo:Nasalislarvatus	Proboscis Monkey	Nasalis larvatus, aka the proboscis monkey, is a primate species endemic to the island of Borneo. Characterized by its large, pendulous nose in males, this arboreal monkey primarily inhabits mangrove forests, riverine, and coastal areas, and is known for its distinct vocalizations and swimming abilities.
foo:Soil	Soil	A dataset describing soil properties from organic and mineral soil across various land uses in Sabah, Malaysia, sampled and measured at the Forest Research Centre Sabah Malaysia.
foo:ManisJavanica	Sunda Pangolin	Sunda pangolin aka Manis Javanica is a mammal distinguished by its protective armor of keratin scales, which cover its body except for its belly and face. Native to Southeast Asia, including Malaysia, Thailand, Indonesia, and Vietnam, this species is adapted to various habitats, ranging from primary and secondary forests to wetlands, mangroves, and grasslands. Characterized by its elongated body, small head, and long, prehensile tail, the Sunda pangolin is primarily nocturnal and has a diet mainly consisting of ants and termites, which it extracts using its long, sticky tongue. It plays a vital role in its ecosystem by controlling insect populations. Manis Javanica is a species critically threatened by poaching and habitat loss. It is one of eight pangolin species, all of which are considered Vulnerable, Endangered, or Critically Endangered according to the IUCN Red List and listed in CITES Appendix I. The Sunda pangolin, critically endangered and the only species found in Malaysia, inhabits Peninsular Malaysia and Malaysian Borneo, including Sabah and Sarawak.
foo:MalayopythonReticulatus	Reticulated Python	Malayopython reticulatus, aka the reticulated python, is a large snake species native to Southeast Asia. Renowned for its impressive length, it is the longest snake in the world, often exceeding 6 meters. It inhabits various environments, including rainforests, woodlands, and plantations, demonstrating adaptability. As a generalist predator, it feeds on many animals, contributing to its ecological significance.
foo:Sensor (Reused from SOSA)	Sensor	Device, agent (including humans), or software (simulation) involved in, or implementing, a Procedure. (e.g., Temperature sensor, humidity sensor, motion sensor). In our model, we have created a unique ID for each sensor based on the platform it is hosted by.
foo:ObservableProperty (Reused from SOSA)	Observable Property	An observable quality (property, characteristic) of a FeatureOfInterest. (e.g., Temperature, humidity, presence)
foo:Observation (Reused from SOSA)	Observation	Act of carrying out an (Observation) Procedure to estimate or calculate a value of a property of a FeatureOfInterest (e.g., Elephant). Observation can be seen as a placeholder that links relevant information together. In our ontology, observation can be considered an ID for each data record.
foo:FeatureOfInterest (Reused from SOSA)	Feature of Interest	The thing whose property is being estimated or calculated in the course of an Observation to arrive at a Result, or whose property is being manipulated by an Actuator, or which is being sampled or transformed in the act of Sampling. In the context of FOO, Soil is the FeatureOfInterest. Most of the sensors are used to observe a property (phenomenon) of a location (e.g., the moisture of soil).
foo:Mammalia (Reused from WO)	Mammalia	The highest class of the subphylum Vertebrata comprising humans and all other animals that nourish their young with milk secreted by mammary glands.
foo:Chordata (Reused from WO)	Chordata	A large phylum of animals that includes the vertebrates together with the sea squirts and lancelets. They are distinguished by the possession of a notochord at some stage during their development.
foo:Elephantidae (Reused from WO)	Elephantidae	ELEPHANTIDAE is a family of bulky mammals (order Proboscidea) comprising the recent elephants and related extinct forms.

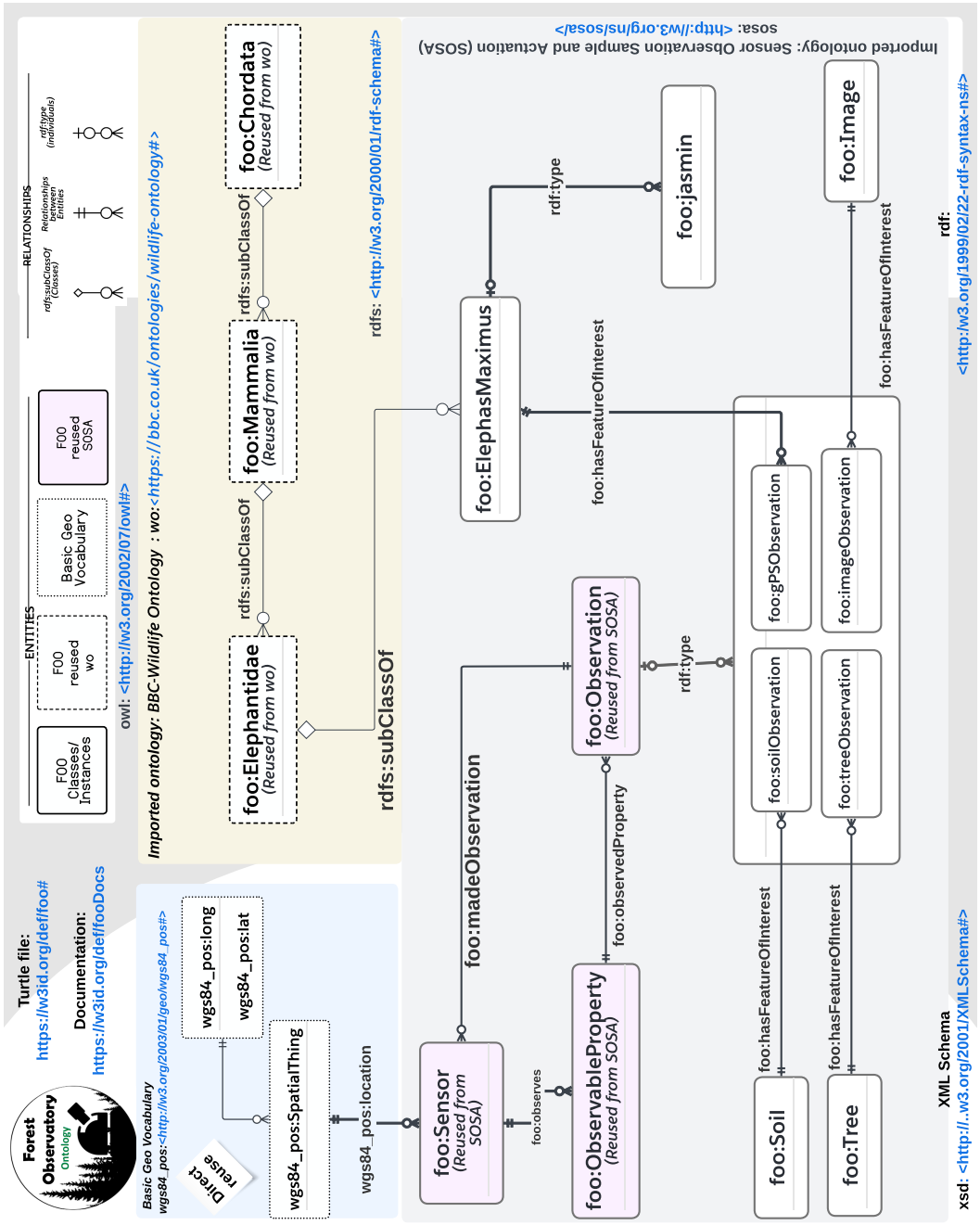


Figure 3.4 Lightweight version of the Forest Observatory Ontology (FOO), main classes, properties and instances.

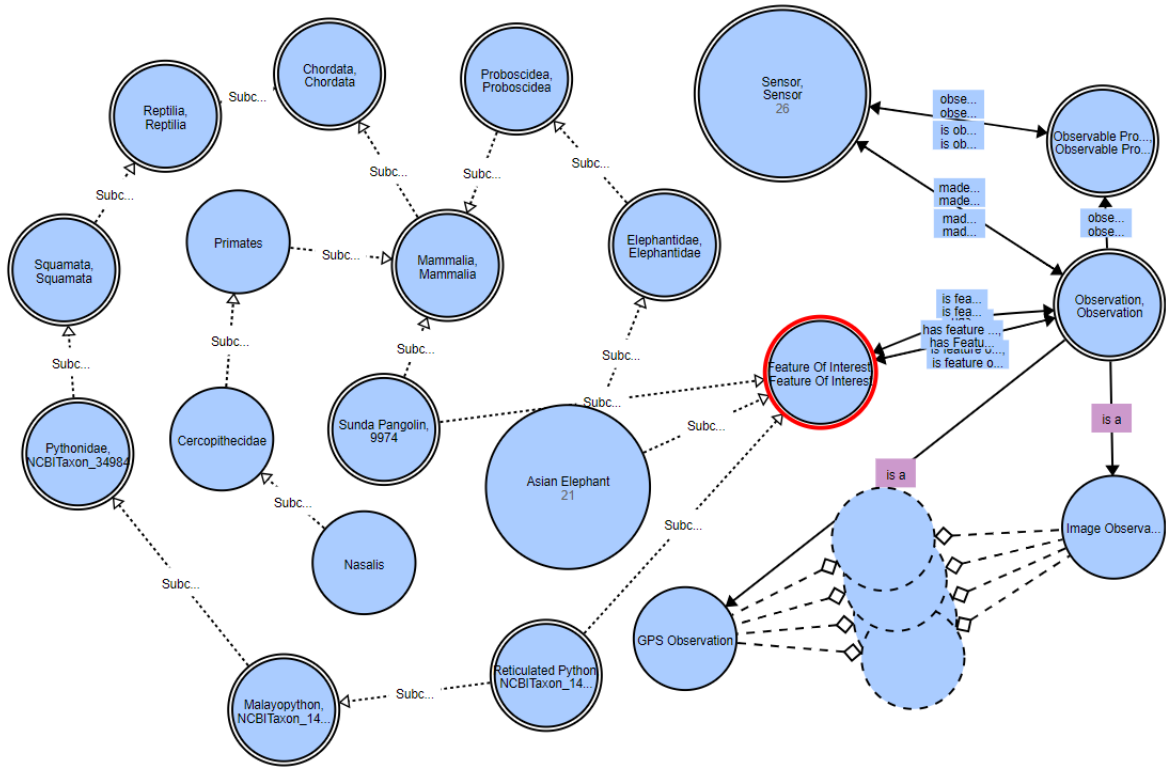


Figure 3.5 Heavyweight version of the Forest Observatory Ontology (FOO), main classes, properties and instances. FOO has 81 classes, 73 properties and 176 instances.

and accessible. The *Interoperable* dimension includes three checks that determine if the vocabulary references pre-existing vocabularies within its metadata annotations, classes, properties, or data properties.

Finally, there are nine checks in the *Reusable* dimension that make sure there is human-readable documentation, provenance metadata, licence information, detailed vocabulary metadata, and that ontology terms are well-described with labels and definitions. FOO passed the test and scored a respected 78%, outperforming SOSA ontology with a score of 67%. Figure 3.6 shows a screenshot of the test results of both SOSA and FOO.

Pellet Evaluation

Pellet, an open-source OWL-DL reasoner, is renowned for its competent performance in identifying conflicting facts in ontologies, making inferences, and responding to SPARQL queries [231]. To evaluate FOO for inconsistencies, I used Protegé's plug-in Pellet version to reason over the ontology. Pellet processed FOO in just 29 ms, calculating the inferences for

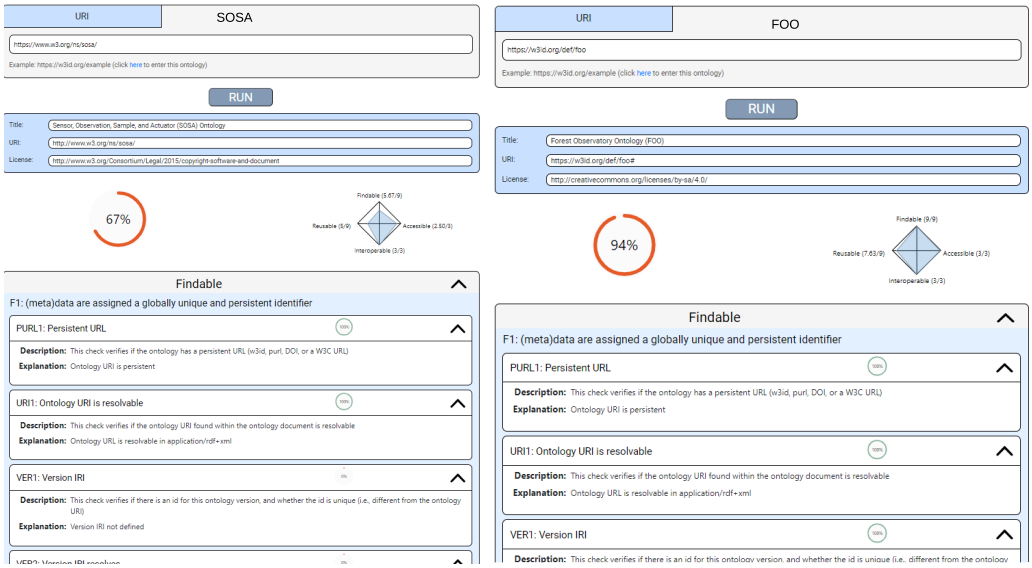


Figure 3.6 FOOPS! Results

the entities' hierarchies and detecting no contradictions. The report generated by the reasoner can be found in FOO's GitHub repository.

Evaluation with SPARQL Queries

To evaluate FOO with SPARQL queries, the structure of the ontology was queried by inspecting its classes, properties, and instances. FOO has 81 classes, 73 properties and 176 instances. SPARQL queries were formulated to explore FOO and subsequently evaluated the performance of each query. The table presented in 3.6 is a summary of the performance metrics for various SPARQL queries used to evaluate FOO. Each row in the table provides a concise description of a specific query's function, such as retrieving all classes, properties, instances, triples, or labels within the ontology. The performance of each query is evaluated based on three key metrics: latency, precision 3.1, and recall 3.2.

Latency, measured in seconds, indicates the time taken to execute the query. As a negative-oriented metric, lower latency signifies faster response times. Precision and recall are used to evaluate the accuracy of query results against expected results (i.e., ground truth) retrieved from FOO beforehand. Precision measures the ratio of relevant instances correctly retrieved by the query to the total instances retrieved, reflecting the accuracy of the results. Recall, on the other hand, measures the ratio of relevant instances retrieved to the total number of relevant instances available, indicating the completeness of the query results.

A score of 1 in both precision and recall was achieved. It means that all retrieved results are relevant (precision) and all relevant results have been retrieved (recall) and that's

justifiable because the ontology and queries were perfectly aligned. This executable notebook shares all the executable codes that queried FOO from its URL.

Table 3.6 Performance for SPARQL Queries Latency in Seconds

Description	SPARQL Query	Latency (s)
Retrieve all classes in the ontology	`SELECT DISTINCT ?class WHERE {?class rdf:type owl:Class .}`	0.0067
Retrieve all properties in the ontology	`SELECT DISTINCT ?property {?property rdf:type owl:ObjectProperty .}`	0.0090
Retrieve all instances of a specific class	`SELECT DISTINCT ?instance {?instance rdf:type foo:Sensor .}`	0.0076
Retrieve labels for all classes	`SELECT DISTINCT ?class ?label {?class rdf:type owl:Class . ?class rdfs:label ?label .}`	0.0091
Retrieve instances with specific properties	`SELECT * {?instance rdf:type foo:Observation ; foo:madeBySensor foo:Jasmin ; foo:hasFeatureOfInterest ?FeatureOfInterest .}`	0.0080
Retrieve all instances and their labels	`SELECT ?instance ?label {?instance rdf:type foo:Sensor . ?instance rdfs:label ?label .}`	0.015
Retrieve instances with their labels and definitions	`SELECT * {?FeatureOfInterest rdf:type foo:FeatureOfInterest; rdfs:label ?label ; skos:definition ?definition .}`	0.013

rdf: <http://w3.org/1999/02/22-rdf-syntax-ns#> owl: <http://w3.org/2002/07/owl#> foo: <https://w3id.org/def/foo#>

The precision metric is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.1}$$

The recall metric is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.2}$$

3.2.4 Ontology Publication and Maintenance

When creating ontologies, it is a common practice to use editors to export them in formats such as Turtle, RDF/XML, and JSON-LD. However, these formats can be complex to understand and use. To address this challenge, researchers can turn to articles or technical reports. However, these sources often prioritise scientific contributions to the detailed

definition of each ontology entity. An alternative solution is to document ontology entities. The semantic web community has developed tools that extract annotation properties from OWL ontologies and generate HTML documentation for classes, properties, and instances. This approach can aid in making ontologies more accessible and understandable [58].

WIZARD for DOCumenting Ontology (WIDOCO) was selected, a tool based on the Live OWL Documentation Environment (LODE), which is utilised within the seven-star linked data model platform [94, 208], to document FOO. WIDOCO enabled us to generate HTML pages that present human-readable and machine-readable visualisations of FOO along with Oops! evaluation. Moreover, OnToology (ontoology.linkeddata.es) [11] was used to secure a persistent identifier for FOO documentation (https://w3id.org/def/foo#), ensuring it can be reliably referenced and accessed over time under the Creative Commons 4.0 International SA (CC BY-SA 4.0) license. I have made FOO and its associated documentation available on FOO's GitHub page to facilitate collaboration and interoperability with other software applications within the research community and for maintenance purposes. Adhering to W3C best practices, I ensured FOO's accessibility in various interoperable formats on the web and deposited it in the BioPortal repository. FOO and its documentation are accessible online through dedicated website (ontology.forest-observatory.cardiff.ac.uk).

3.3 Forest Ontology Observatory Data Store (FooDS)

This section describes FooDS as an ontology-based knowledge graph built with four distinct datasets. FooDS enabled the representation and integration of diverse wildlife data sources in a unified manner. Figure 3.8 shows the relationships between the proposed ontology (FOO) and wildlife knowledge graphs. To transform four wildlife datasets —encompassing soil data, vegetation and site habitats, GPS collar data, and trail camera images into knowledge graphs, Matey web user interface (rml.io/yarrrml/matey), powered by YARRRML (Yet Another RDF Rules Language) [120, 251] was employed. YARRRML (rml.io/yarrrml) specifies a set of prefixes to create namespaces and offers mapping rules to generate RDF triples from the data sources. Figure 3.8 explains how FOODS was generated. In addition, modular pipelines (i.e., Python scripts) were developed to manage large data volumes, ensuring data serialisation with names or schemas that align with those defined in FOO. Resources including the ontology and code, are available on the proposed website (ontology.forest-observatory.cardiff.ac.uk).

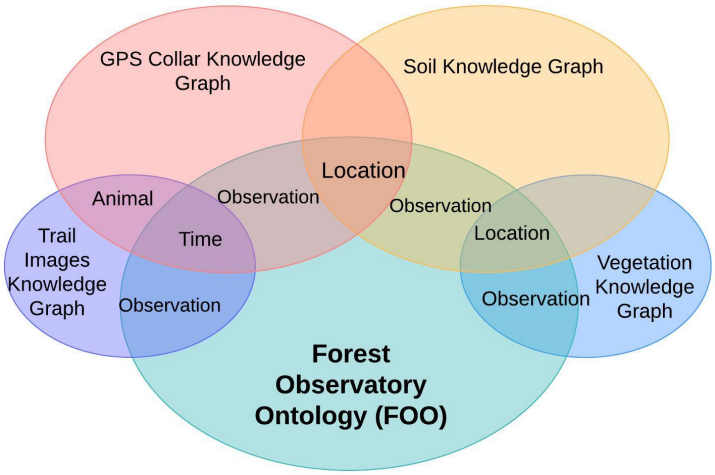


Figure 3.7 Main related concepts between FOO and the proposed knowledge graphs

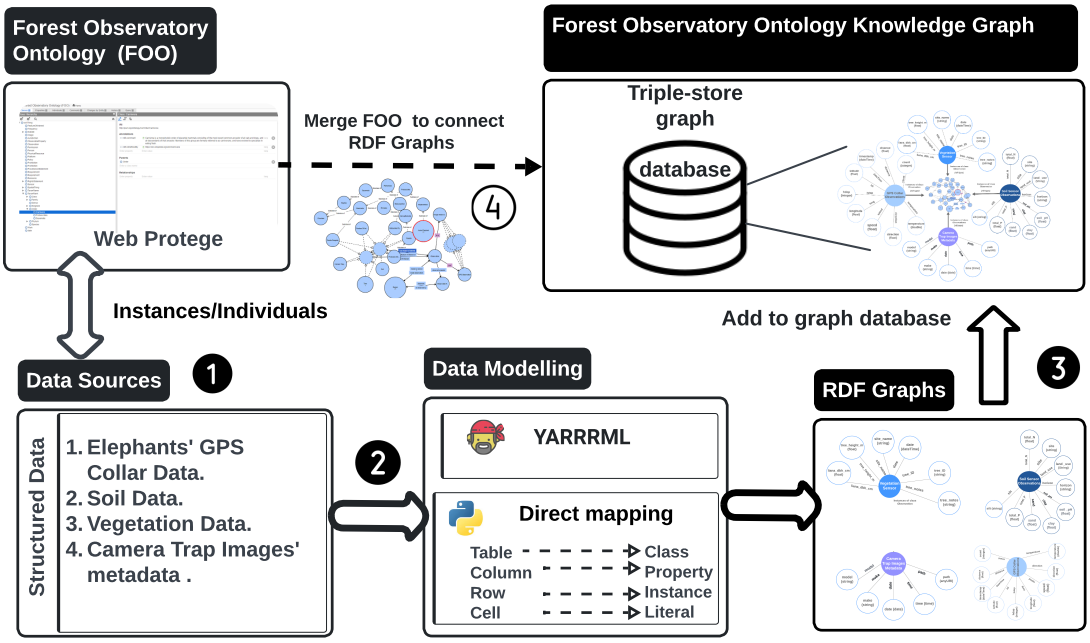


Figure 3.8 The proposed ontology-based Knowledge graph construction approach

3.3.1 Soil RDF Graph

Based on the experience drawn from ontology development, a decision was made to outsource the soil data. The nominated dataset contained characteristics and nutrient content for logged and unlogged tropical forests in Sabah, Malaysia. Soil properties were obtained using buried ion-exchange membranes, and nutrient levels were measured. These data were made possible by the BALI collaboration, which was funded by the UK's Natural Environment Research

Council (NERC) [79]. YARRRML (Yet Another RDF Rules Language) syntax was used to define RDF mappings in a human-readable format. YARRRML specifies prefixes to define namespaces and shorthand notations for Uniform Resource Identifiers (URIs). The mappings represented in listings 3.1, 3.2 defined the rules for the "Observation" entity. These mappings specified the data source as "soil.csv csv" and mapped the observation properties using the "s" subject template, which combines the foo namespace with the value of the "Identifier" column. The po (predicate, object) mapped section lists the properties and their corresponding values for observation. Figure 3.9 shows the classes and instances distribution for the soil knowledge graph. Meanwhile, Table 3.7 provides a descriptive analysis of the modelled data.

Listing 3.1 Soil data prefixes

```
foo: "http://w3id.org/def/foo#"
xsd: "http://w3.org/2001/XMLSchema#"
```

Listing 3.2 Soil data YARRRML

```
mappings:
  soil:
  sources:
    - ['soil.csv~csv']
  s: foo:$(Identifier)
  po:
    - [a, foo:Observation]
    - [foo:Site, $(Site)]
    - [foo:Land_Use, $(Land_Use)]
    - [foo:Plot_Name, $(Plot_Name)]
    - [foo:Subplot, $(Subplot)]
    - [foo:Horizon, $(Horizon)]
```

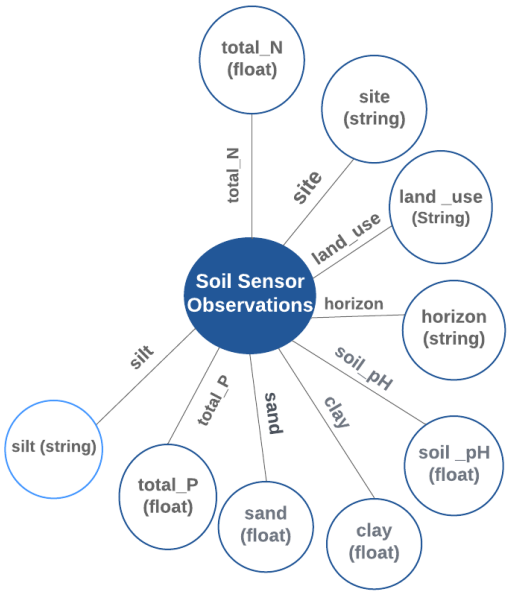


Figure 3.9 Soil Knowledge Graph

Table 3.7 Soil data descriptive analysis

	Subplot	Soil_Moist.	Horizon_Depth	Bulk_Density	Soil_pH	total_C	total_N	Inorganic_P	C N	C:P
Count	222	222	222	222	222	222	222	222	222	222
Mean	13.18	26.71	3.74	0.69	5.59	6.04	0.39	33.97	15.04	0.27
STD	7.93	9.37	1.91	0.28	0.85	4.61	0.21	51.56	3.52	0.16
MIN	1.00	7.62	0.20	0.17	3.22	0.83	0.09	3.77	6.65	0.01
25%	6.00	20.56	2.35	0.49	4.92	3.74	0.27	14.34	12.96	0.17
50%	12.00	26.43	3.50	0.66	5.58	4.94	0.34	20.49	14.46	0.24
75%	20.00	31.91	5.00	0.86	6.32	6.33	0.43	32.91	16.47	0.33
MAX	25.00	65.10	9.50	1.84	7.42	33.45	1.49	571.25	39.59	0.89

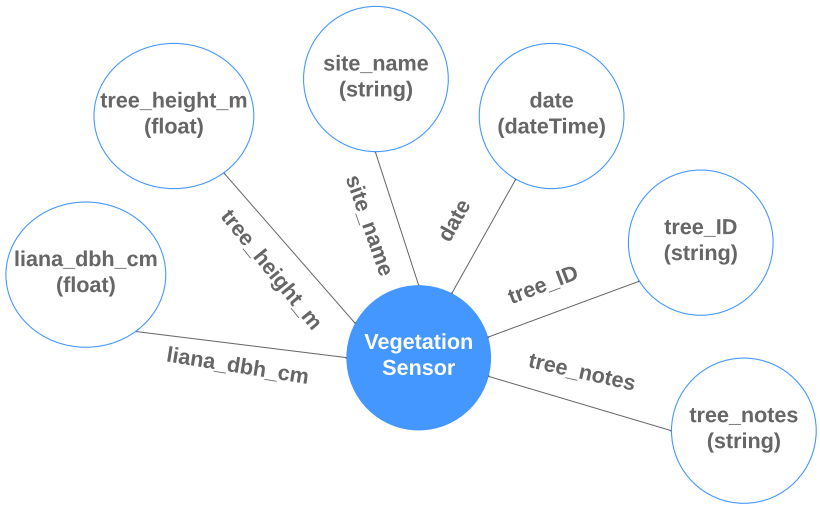


Figure 3.10 Vegetation RDF Graph

3.3.2 Vegetation RDF Graph

Another nominated Open data contained records of plants from 49 plots in Sabah, Malaysian Borneo, spanning 14 fragmented forest areas and four continuous forest sites. The vegetation data collected from two to three sites in each of the 18 locations included information on living plants and dead trees. The data contained plant properties, forest structure measures, and forest fragmentation metrics in the surrounding landscape of the plots. The primary objectives of collecting these data were to support research focused on (i) understanding the factors driving the spread of exotic plant species in fragmented forest areas and (ii) evaluating the effectiveness of conservation set-asides in palm oil plantations to preserve carbon storage and plant diversity [80]. Figures 3.10 and Table 3.8 illustrate the vegetation knowledge graph modelling and its descriptive data analysis, respectively. The RDF mappings were modelled in a human-readable format using YARRRML (Yet Another RDF Rules Language) syntax. These mappings defined rules for the "Observation" entity by specifying the data source as "veg.csv csv" and mapped observation properties through the "s" subject template,

Table 3.8 Lianas data descriptive analysis

	Plot_no	Tree_indv_no	Tree_dbh_cm	Tree_ht_m	Tree_N_lianas	Liana_dbh_cm	Subplt_radi._m	ID1
Count	3070.00	3070.00	3070.00	1103.00	3070.00	3070.00	3070.00	3070.00
Mean	2.01	1984.24	30.09	20.05	5.74	3.91	25.51	1535.50
STD	0.83	1132.74	17.66	10.73	4.65	2.27	4.97	886.38
MIN	1.00	2.00	10.00	3.00	1.00	2.00	20.00	1.00
25%	1.00	1028.00	17.50	12.00	3.00	2.40	20.00	768.25
50%	2.00	2094.50	26.30	17.00	4.00	3.20	30.00	1535.50
75%	3.00	3022.00	37.00	25.00	7.00	4.60	30.00	2302.75
MAX	3.00	3895.00	140.00	60.00	31.00	21.80	30.00	3070.00

which merges the FOO namespace with values from the "Site_name" column. The po (predicate, object) map section enumerates the properties and their corresponding values for the observation. This mapping approach was similarly applied to the GPS collar and trail image data.

3.3.3 GPS Collar RDF Graph

The GPS collar datasets were acquired from the Danau Girang Field Centre (DGFC). These sets included data from GPS collars fitted on twenty-two adult Asian elephants, encompassing 14 females and eight males. The fitting process involved a collaborative effort among researchers, trackers, and a wildlife veterinarian. Supplied by Africa Wildlife Tracking, the collars weighed approximately 14 kg and were equipped with a Global Positioning System (GPS) receiver and a Very High Frequency (VHF) transmitter. Between 2012 and 2018, these devices systematically recorded data on time, location, and temperature, among other variables, at two-hour intervals, as detailed in Table 4.4 [163, 81]. Figure 3.11 shows the distribution of classes and instances within the collar knowledge graph. Owing to the sensitive nature of the data and the risk of poaching, it will not be made publicly accessible, prioritising the protection of these endangered species [163, 81].

3.3.4 Camera Trap Images RDF Graph

A dataset containing 1000 images of Asian elephants was modelled. Before their transformation into RDF graphs, the images' metadata were extracted and stored as CSV files. The RDF dataset includes unique paths that point to image locations on a protected cloud server. Figure 3.12 and illustrates the results of the semantic modelling and the data entities, respectively.

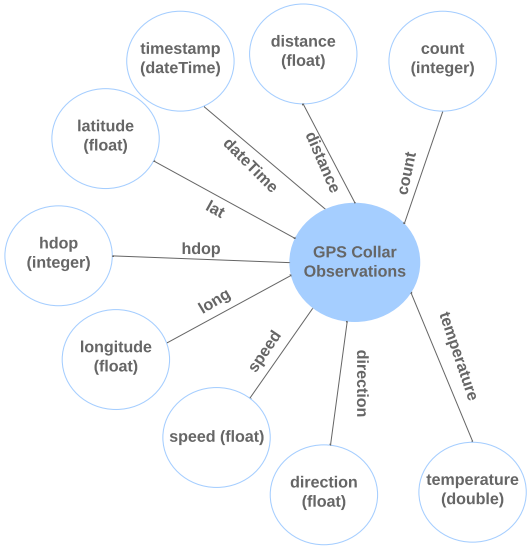


Figure 3.11 GPS Collar RDF graph

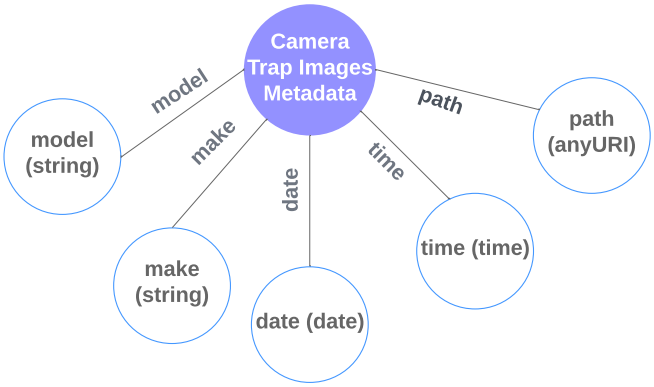


Figure 3.12 Camera Trap Images' metadata knowledge graph conceptual model

3.3.5 Semantic Data Integration

Figure 3.13 exemplifies how FOO integrates and links the four datasets of interest (RDF graphs) to form FooDS. The overall architecture and elements of the FooDS are depicted in Figure 3.14. In this system, wildlife data collected from various research activities are managed by a data manager, who assigns each dataset to an RDF graph using its specific mapper code. These RDF graphs and the Forest Observatory Ontology (FOO) are stored together in a unified database known as a triple store. Creating a knowledge graph involves mapping data from the source schema—the schema of the original data source—to the target schema—the schema of the knowledge graph. This target schema is represented here by RDF,

which is a structured format governed by a vocabulary or ontology [249]. FOO interlinks these diverse datasets with a shared URL for each dataset, referencing their conversion into RDF format. Once these RDF formatted datasets are merged within the same triple store, they inherently connect and form an ontology-based knowledge graph. The ontology-based knowledge graph(s) are stored and published as URLs, which can be accessed and parsed using libraries such as RDFLib in Python environments like Colab. This accessibility allows easy integration with existing data analysis tools and enhances collaborative opportunities across the scientific community.

This query language enables authorised users— wildlife researchers, data scientists, and developers— to access and manipulate the graphs. FooDS provides a robust foundation for integrating AI technologies to enhance the capabilities of intelligent systems. The formal, logic-based representation of knowledge in the knowledge graph enables the application of semantic reasoning techniques, such as rule-based inference and probabilistic reasoning, to derive new insights and make inferences. Natural language processing can extract entities, relationships, and attributes from unstructured text and populate the knowledge graph. At the same time, the graph structure can also improve NLP tasks by providing valuable contextual information. Machine learning models can be trained on the structured data in the knowledge graph to perform classification, prediction, and recommendation, with the relational features enhancing the accuracy and interpretability of these AI models. The flexible knowledge representation in the graph also enables deep learning techniques to learn vector representations of entities and relationships, improving reasoning and inference. Furthermore, the semantic nature of knowledge graphs can help make AI systems more transparent and explainable by tracing the reasoning behind outputs using encoded relationships and logical rules. Knowledge graphs can integrate diverse data sources, and AI techniques like entity resolution and data fusion can be applied to maintain data quality and consistency.

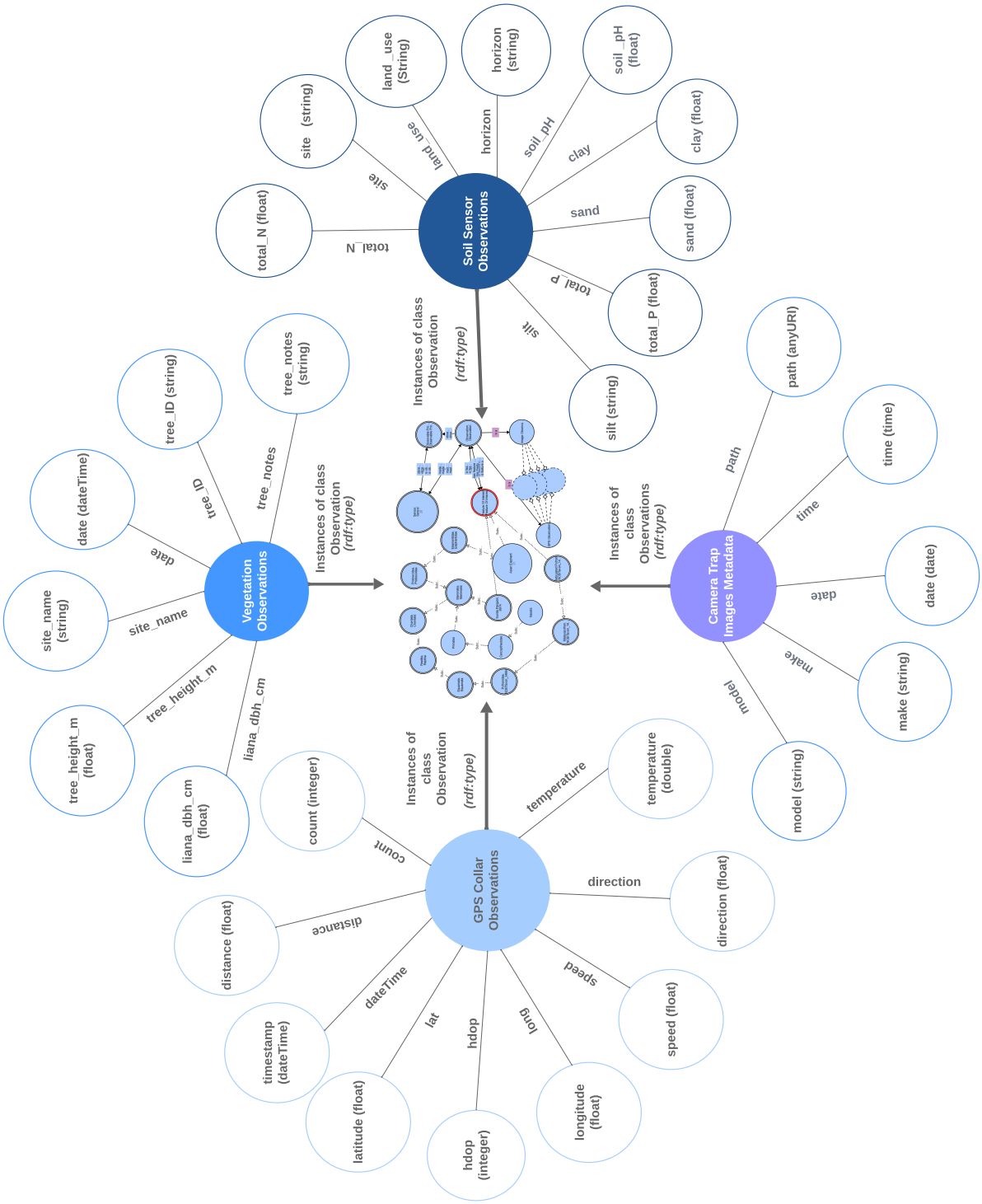


Figure 3.13 FOO is populated with the four heterogeneous datasets (above) transformed into RDF graphs, referencing FOO URI (w3id.org/def/foo) to form FOODS.

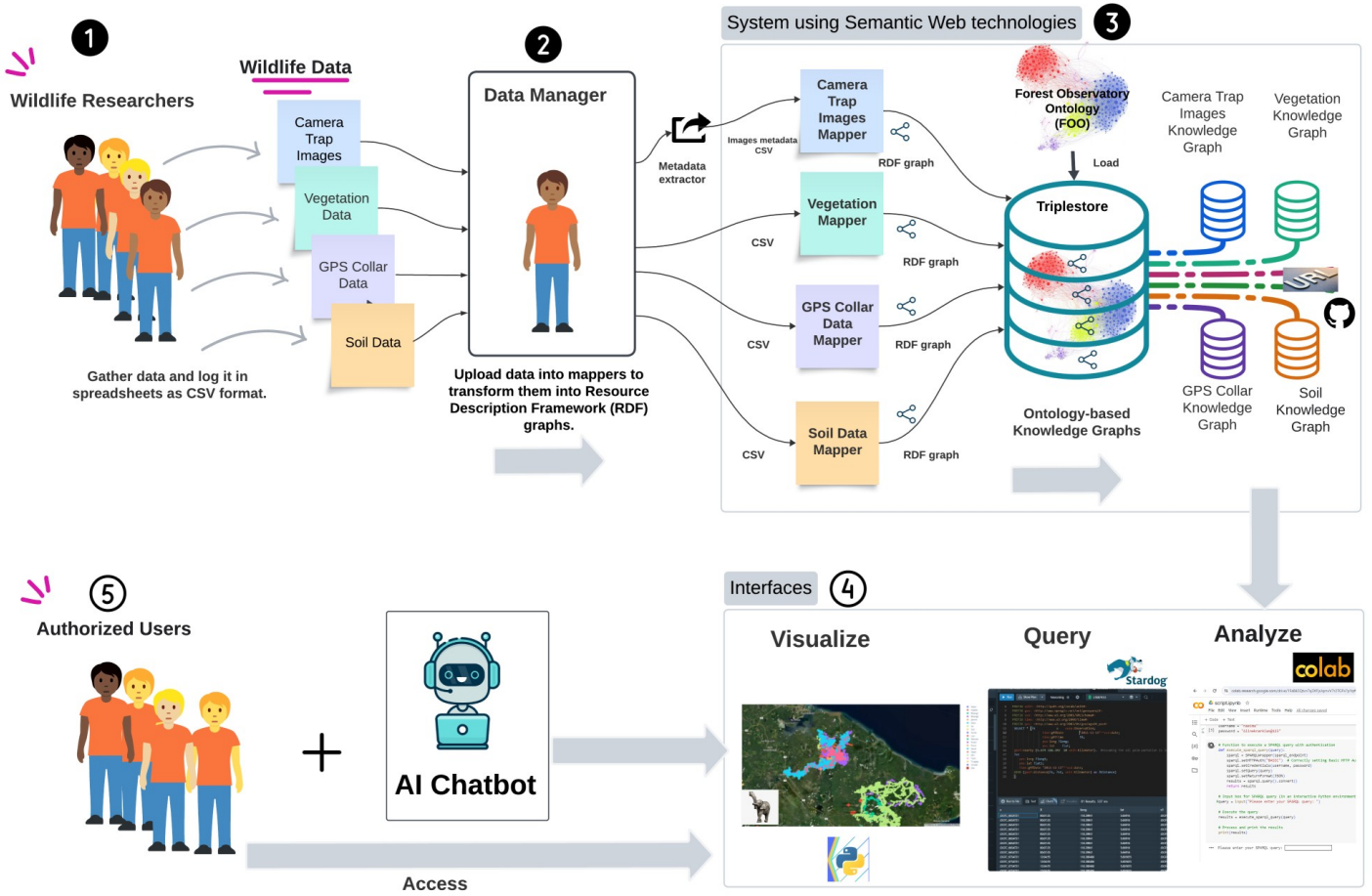


Figure 3.14 This figure shows the management of wildlife research data through RDF graph assignment by a data manager, with RDF graphs and the ontology FOO stored in a unified database (triple Store). It highlights the process of knowledge graph generation as a mapping between source and target schemas, focusing on RDF for the target schema. The division of ontology-based knowledge graphs into four distinct graphs for different data types (e.g., soil knowledge graph), stored separately, on the same platform, and published online.

Participants	Q1.Confidence	Q2.Usefulness	Q3.Ease of use	Q4.Performance	Q5.Saves Time	Q6. UI Clarity	Q7. Tech Support	Q8. Visual.	Q9.Data Handle	Education	Occupation	Experience
P1	4	4	5	5	4	5	4	5	5	Master's	Researcher	4 to 10
P2	4	4	2	3	5	4	3	4	5	Master's	Researcher	4 to 10
P3	4	5	3	4	4	4	1	5	4	Master's	Researcher	4 to 10
P4	5	5	4	5	5	4	2	5	5	Doctorate	Researcher	4 to 10
P5	5	5	4	5	5	4	3	5	5	Master's	Data Scientist	4 to 10
P6	5	5	5	5	5	5	4	4	5	Master's	Researcher	1 to 3
P7	5	4	4	5	5	4	3	5	5	Bachelor's	Data Manager	11 to 20
P8	5	4	3	5	4	4	3	5	5	Doctorate	Researcher	4 to 10
P9	3	3	3	3	4	3	3	4	3	Doctorate	Conservation Biologist	4 to 10

Table 3.9 Usability study results. (1= Strongly Disagree), (2 = Disagree), (3= Neutral), (4= Agree), (5= Strongly Agree). Q1. I feel confident in the tool's ability to merge and manage data from multiple sources. Q2. The tool is useful in answering questions from different data sets. Q3. Learning to use the data integration tool can be easy. Q4. The tool's performance (speed, stability) meets my expectations. Q5. Integrating data using this tool saves me time. Q6. The user interface of the data integration tool is clear and understandable. Q7. I require technical support frequently when using this data integration tool. Q8. The tool provides clear visualisation of different animals' movements. Q9. I am satisfied with how the data integration tool handles complex data sets.

3.4 FooDS Evaluation

This section assesses the ontology-based knowledge graphs using a task-based approach. To evaluate the usefulness of FOODS, I conducted a *usability study involving nine domain experts* to assess system performance; six of these experts are related to DGFC and participated in the discussion groups during the ontology requirement gathering phase. I then discuss three use cases—3.4.2, 3.4.3, and 3.4.4—derived from the requirements outlined in Section 3.2.1. The use cases created based on data collected from the discussion groups. The reasoning inferred events that are not directly expressed in the data. For example, hazardous areas that can put the elephants at risk. I conducted an in-depth evaluation of the third use case (3.4.4), as it encapsulates a real-life scenario. This evaluation highlighted the primary benefits of FOODS, particularly by extracting several Competency Questions (CQs) from use cases.

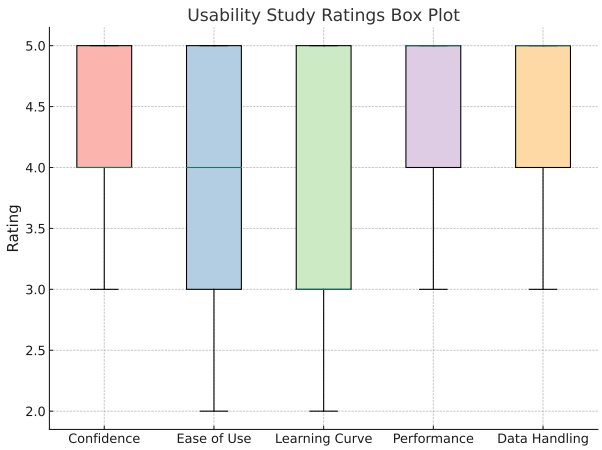


Figure 3.15 Box plot representation of usability ratings across five key metrics (Confidence, Ease of Use, Learning Curve, Performance, Data Handling) as evaluated by participants in a usability study. Each box indicates the interquartile range (IQR) with the median highlighted, encapsulating the central tendency and dispersion of ratings, thus providing insights into the tool's perceived effectiveness and user satisfaction.

3.4.1 Domain Experts Evaluation

The usability study for testing the ontology-based knowledge graphs was judged by the presence of a conservation biologist among the participants. Participants were provided presentations about the dashboard and how to query and analyse the knowledge graphs. Responses were quantified on a Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) across several aspects such as confidence in the tool, its usefulness, ease of learning, performance, time efficiency, UI clarity, need for technical support, visualisation quality, and data handling satisfaction. Table 3.9 shows participant feedback on our proposed ontology-based knowledge graphs across various dimensions. I analysed and visualised the results, reducing the dimensions to confidence, ease of use, learning curve, performance, and data handling capabilities for simplicity (Figure 3.15).

3.4.2 Use Case 1: Elephants spending time together

Elephants, as mammals, maintain connections with their families and interact with elephants from other herds. They engage in activities such as travelling, foraging, and socialising. Their interactions in the wild can be complex and vary based on factors such as age, sex, and familial ties. Researchers have employed GPS collars and motion-activated trail cameras to observe elephant behaviour and track their movements in their natural habitat. Understanding the migration patterns of elephants in Sabah forest is vital for shaping forest management

strategies. It helps identify key conservation habitats, reduces human-elephant conflicts, and guides the allocation of resources, such as deploying rangers and fitting motion-activated cameras. These patterns reveal where elephants move and find essential resources, allowing for focused conservation initiatives, strategies to minimise conflicts and practical use of anti-poaching resources. Moreover, it enables researchers to deduce when different elephants spend time together. For example, observing two elephants travelling or foraging in the same geographic area could indicate their social interactions. Access to data from GPS collars, soil sensors, and camera traps collectively aids researchers in understanding elephant social dynamics and migration patterns. To illustrate this concept, the SPARQL query in .1 (CQ19) was formulated to find out which elephants met.

3.4.3 Use Case 2: Salt licks locations

Salt or mineral licks are natural deposits of salts and minerals that animals consume as essential nutrients. In Sabah, several protected areas, such as the Danum Valley Conservation Area, Tabin Wildlife Reserve, and Maliau Basin Conservation Area, harbour salt licks that attract elephants and other wildlife species. The exact locations of these salt licks may be kept confidential to prevent disturbance or exploitation of wildlife. Elephants can obtain vital minerals and nutrients from salt licks, which may not be readily available in their regular diet. However, excessive use of salt licks can lead to detrimental effects such as overgrazing and soil erosion, harming the surrounding ecosystem. Having a tool that provides access to curated and semantically integrated data about elephant GPS locations and information about salt lick areas, including soil conditions and vegetation, can empower wildlife conservationists to make informed decisions to protect the natural habitats and resources of wildlife. The SPARQL query listed in Listing allows users to select elephants observed in the Danum Valley and gather information about salt licks in the area. In this use case, the sensor observations include the elephant's name, which can be selected in the query .1 (CQ72).

3.4.4 Use Case 3: Rescuing the injured elephant

Numerous elephants in the Kinabatangan region are equipped with GPS collars to track and monitor their movements. These collars are named after the elephants to which they are attached (e.g., Jasmin, Seri, Sandi, etc.). Bioscientists regularly access and visualise real-time data from the collars and store historical data for later analysis. During one such analysis, a chief scientist observed an unusual pattern in the GPS data for elephant Jasmin; the observations were repeated at the exact location for two days. Consequently, a wildlife

officer was dispatched to check Jasmin, leading to the discovery that a snare near an oil palm plantation injured the elephant. The officer promptly notified the manager, who contacted a veterinarian to rescue Jasmin. The proposed solution involves designating predetermined geographical boundaries. If an animal is found to have crossed these boundaries, it should be treated as a potential danger that may necessitate intervention. Due to availability, this system currently uses historical data. However, at the reproduction level, our system can stream real-time sensor data into our pipeline codes for transformation into RDF. I developed use cases to evaluate the proposed knowledge graphs, focusing on three main tasks: integrating heterogeneous wildlife data from various sources, providing precise and immediate data retrieval, and demonstrating the ability to deduce novel information through reasoning techniques. These use cases served as benchmarks to assess the knowledge graphs' performance and effectiveness in achieving these objectives. I derived four Competency Questions (CQs) based on the third use case, Listing , which depicts a real-life scenario. These questions were formulated to SPARQL queries in .1 (CQ4, CQ17, CQ54, CQ92) to evaluate the effectiveness of the knowledge graphs in providing accurate answers.

- CQ1: What are elephant Jasmin's observations between 2012-02-07 and 2012-02-15?
- CQ2: When did elephant Jasmin go near the plantation on 2012-02-07?
- CQ3: What are the soil metrics near the elephant Jasmin?
- CQ4: What are the other elephants near the palm oil plantations?

CQ1 investigates how GPS collar information can be merged with camera-trap images to authenticate elephant identities and assess the urgency of incidents. CQ2 aims to elucidate elephant behaviours, such as foraging and socialising. CQ3 delves into the significance of soil conditions near elephant locations, which can influence their speed and movement, for instance, by causing their legs to become stuck in mud if the soil is wet. CQ4 leverages the reasoning capabilities of semantic web technologies to introduce assertive rules into data. In artificial intelligence and knowledge representation, reasoning is essential for enabling systems to draw conclusions from available data and established rules. Reasoning approaches are commonly divided into two primary types: deductive reasoning, which derives conclusions based on general premises, and rule-based reasoning, which relies on predefined "if-then" rules to make specific inferences in given contexts. For instance, a logical rule might stipulate that if a snare injures one elephant, the other nearby elephants could also be at risk. I used Federated SPARQL queries to interrogate the knowledge graphs, enabling us to retrieve answers from any integrated data source within the FOO. Responses

Table 3.10 Statistics for the four Competency Questions (CQs) query responses in milliseconds (ms)

	CQ1		CQ2		CQ3		CQ4	
Description	NO-Rule	SWRL	NO-Rule	SWRL	No-Rule	SWRL	No-Rule	SWRL
Mean	54.60	43.04	88.98	70.64	262.30	207.76	2083.42	1646.26
Std	7.84	21.66	25.94	37.27	47.14	102.57	47.17	791.33
Variance	61.46	469.15	972.88	1389.05	2222	10520.6	2225.00	626.203
Minimum	43.00	-6.00	60.00	-8.00	205.00	-25.00	2029	-208.00
Maximum	77.00	89.00	172.00	173.00	390.00	427.00	2246	3064
Shapiro-Wilk test p-value (0.05)	0.001	0.755	3.20-6	0.367	5.916	0.405	7074-06	0.241
Mann-Whitney U test p-value (0.05)	0.001		0.014		0.006		0.001	
Count	50		50		50		50	

were obtained from FooDS, including data from the GPS collar and soil datasets incorporated into FOO. SPARQL queries are available in .1.

3.4.5 Results

An experiment was conducted to assess responses to four Competency Questions (CQs) based on correctness, completeness, and speed. Using the Stardog Studio knowledge graph platform (cloud.stardog.com), the queries were executed 50 times each without reasoning, with a 3-second interval between executions, and response times were recorded. The experiment was then repeated with reasoning enabled by activating the reasoning option and incorporating Semantic Wb Rule Language (SWRL) into the triple store. The SWRL rule stipulated that if the distance between an elephant and an oil palm plantation is less than 5 km, it constitutes a hazard. Figure 3.16 illustrates the response times for each query, showing average times of 54.7 ms, 87.29 ms, 259 ms, and 2080 ms for CQ1, CQ2, CQ3, and CQ4, respectively. The results were accurate, with CQ4 demonstrating the longest response time due to the large volume of requested information. Similarly, CQ3 had longer response times compared to CQ1 and CQ2 because of the necessity to link disparate databases. Although response times were faster after enabling reasoning, the correctness and accuracy of the query results remained consistent. To analyse the data, the Shapiro-Wilk test for normality and the non-parametric Mann-Whitney U test were applied to compare the two sets of independent responses. The Shapiro-Wilk test indicated that query responses without rules were not normally distributed, whereas a normal distribution was observed for the SWRL responses. Given the non-normal distribution of the rule-free query responses, the Mann-Whitney U test was employed and revealed a significant difference between responses before and after enabling reasoning or incorporating SWRL. Table 3.10 shows the statistics of the CQs response time, while complete documentation for this evaluation is available at *https://github.com/Naeima/Knowledge-Graphs-Evaluations.git*.

Listing 3.3: Shapiro Wilk test

```
Hypotheses
H0: Samples are normally distributed
H1: Samples are not normally distributed
p-value cutoff = 0.05  if p > value:
Retain H0 - Samples are normally distributed.
Else:
Reject H0 - Samples are not normally distributed.
```

Listing 3.4: Mann-Whitney U test

```
Hypotheses
H0: Samples median are equal
H1 : Samples median are not equal
p-value cutoff = 0.05  if p > value:
Retain H0 - the medians are equal.
Else:
Reject H0 - the medians are not equal.
```

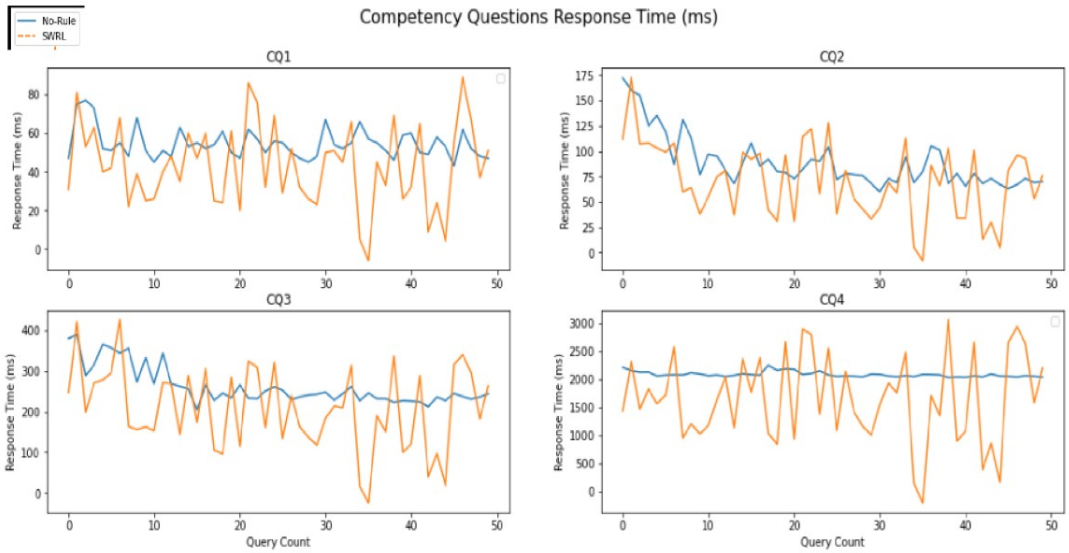


Figure 3.16 Response time for the four Competency Questions (CQs)

3.4.6 Discussion

Forest Observatory Ontology Data Store (FooDS), the ontology-based knowledge graph, was generally well-received by domain experts, with high scores in confidence, usefulness, performance, time-saving, UI clarity, visualisation, and data handling across participants with diverse educational backgrounds and occupations, mainly researchers and data specialists. The ontology-based knowledge graphs' ease of use and the need for technical support received more varied responses, indicating areas where few end-users might require prior indication training. The educational backgrounds and occupations of the participants suggest that the knowledge graphs are relevant to academic and professional research, especially in fields requiring data integration and analysis. FooDS answered 83 Competency Questions (CQs) correctly from 105 CQs and addressed three specific use cases. These 105 CQs were formulated from the data collected during the requirement phase (ethnography, semi-structured Interviews and discussion groups) for example, a SPARQL query could be generic to address elephants while the CQs address specific elephant and specific event. There is

an overlap but I suggest that CQ2 capture the high level of knowledge while use cases CQs are more tailored to solve the task in hand. Use cases CQs are extracted from the general CQs and tunned to address the case. Applying reasoning, particularly through SWRL rules, optimised the inferred new knowledge from data and accelerated the query response times whilst maintaining accuracy. Through thorough statistical analyses of the results from the third use case, it was demonstrated that querying FooDS, stored in a triple store and enriched with SWRL reasoning rules outperforms querying a rule-free one, particularly in speed and efficiency, thereby expediting the search and discovery processes. Challenges related to data sensitivity and scarcity were encountered, particularly concerning the ethical implications of GPS collar data for elephants. This emphasises the need to balance data utility with conservation ethics. The limited scope of the data is compounded by the lack of collared status for many elephants and difficulties in data collection owing to environmental factors. In addition, integrating real-time data into FOO poses distinct challenges, mainly because of the instability of data generation and connectivity issues. Although this framework relies on historical data, I recognise its potential for integrating real-time sensor data streams. This can be achieved using IoT devices, such as Arduino or Raspberry Pi boards, and protocols like MQTT or WebSockets for seamless data transmission. Theoretically, embedding logic into these devices for continuous sensor data collection and converting the data into RDF format would enable real-time data streaming. However, this enhancement, although feasible, is beyond the scope of the current project and is a direction for future development. The need for real-time data is beneficial in the third use case scenario, such as the prompt rescue of injured elephants, where reliance on historical data hinders swift responses to emergencies. The integration of real-time data can facilitate immediate action to aid wildlife injuries. I concede that, in its present configuration, FOO does not offer benefits for the third scenario involving the rescue of an injured elephant unless real-time or near real-time data are incorporated. The proposed framework, centred on defining domain-specific ontologies followed by the data population to generate ontology-based knowledge graphs, offers a flexible and replicable method across various domains. This approach is applied in healthcare [125], smart cities for urban planning [146], finance for market predictions [280], cultural heritage for connecting historical dots [67], and education for personalised learning solutions [47]. This methodology not only aids in structuring domain knowledge but also facilitates the extraction of actionable insights, demonstrating its broad applicability and potential to revolutionise knowledge representation and decision-making across diverse fields.

3.5 Summary

The first part of this chapter discussed the role of Forest Observatories in enhancing the understanding of wildlife dynamics by integrating heterogeneous wildlife data and breaking down data silos. It introduced the Forest Observatory Ontology (FOO) and its knowledge graphs, supported by a resource website (ontology.forest-observatory.cardiff.ac.uk), a digital book, executable notebooks for SPARQL queries, and chatbot (developed as a proof of concept and not formally evaluated) to prove that the knowledge graphs can be accessed by non-technical users. Following this, four knowledge graphs (soil, vegetation, GPS collar, and camera trap image metadata), their semantic modelling and ontology integration processes were proposed, resulting in the Forest Ontology Observatory Data Store (FooDS). In addition, FooDS was evaluated S domain expert feedback and applied use cases. The subsequent chapter used FooDS for deep learning to support bioscientists and conservationists in predicting poaching.

Chapter 4

Leveraging FooDS for Predicting Wildlife Poaching

This chapter addresses the third research question (RQ3): *Can prediction models be developed to predict poaching crimes by using the developed 'Linked Data Store'?*

This chapter draws upon linked data from the previously developed Forest Observatory Ontology Data Store (FooDS) to build predictive models based on deep learning algorithms. These models are designed to forecast the future geo-locations of elephants and assess the likelihood of poaching incidents, supported by rule-based semantic reasoning. The resulting application introduces a hybrid predictive model that leverages heterogeneous wildlife data drawn from an ontology-based knowledge graph. The novelty of this approach lies in its capacity to harmonise deep learning with ontology-based knowledge graph's reasoning to generate accurate, actionable predictions.

4.1 Introduction

Habitat loss, human-elephant conflict, and poaching threaten Bornean elephants (Elephas maximus) [81]. Despite global anti-poaching efforts, the illegal ivory trade continues to drive poaching, reducing the population to fewer than 1,500 [106, 3, 42]. In Sabah, Malaysian Borneo, over 200 elephants died between 2010 and 2021, many through poisoning near oil palm plantations [191, 64]. High-profile incidents, such as the 2013 poisoning of 14 pygmy elephants, highlight the escalating conflict between expanding agriculture and wildlife conservation. Other species, including Bornean orangutans (Pongo pygmaeus), proboscis monkeys (Nasalis larvatus), and Sunda pangolins (Manis javanica), face similar threats from habitat loss and animal trafficking for illegal trade [64]. Poaching also endangers human

lives, with rangers facing armed poachers [180]. Limited and inconsistent poaching data complicate reliable predictions [270], prompting researchers to leverage environmental data for insights. For example, GPS sensors used by the World Wildlife Fund in Sabah help monitor elephant behaviour and mitigate human-elephant conflicts. Advanced machine learning models built on GPS and environmental data can predict wildlife movements, assisting targeted anti-poaching efforts.

This chapter addresses the third research question (RQ3): *Can prediction models be developed to predict poaching crimes by using the developed 'Linked Data Store'?*

To address RQ3, *PoachNet* is presented, a predictive tool designed integrate wildlife data with advanced algorithms. PoachNet employs deep learning with FooDS creating a dynamic and hybrid model for poaching prediction. Elephant GPS observations are processed through a sequential neural network to predict geo-locations, which are semantically modelled and incorporated into FooDS. Semantic Web Rule Language (SWRL) asserts poaching rules based on events not explicitly expressed in FooDS. PoachNet's performance was benchmarked against state-of-the-art methods and demonstrated higher accuracy, consistently outperforming them.

PoachNet allows users to interact with the data through directed queries, enabling the retrieval of granular wildlife data from knowledge graphs either stored locally in a triple-store or accessed online via Uniform Resource Identifiers (URIs). Once extracted, the data are processed through a deep learning model that predicts an elephant's geo-location (latitude and longitude) based on GPS collar data, whilst semantic reasoning assesses the likelihood of poaching in specific areas.

In addressing RQ3, this application achieves the third core contribution (C3) of this research: developing a solution for predicting poaching threatening a Bornean elephant (Elephas Maximus) by using data extracted from FooDS. This contribution assists bioscientists and conservationists by enhancing poaching prediction capabilities through modular, scalable predictive models enriched with semantic data.

4.2 Elephant GPS Collar Knowledge Graph RDF

Sourced from Danau Girang Field Centre [163], Table 4.1 describes 9168 observations from Global Positioning System (GPS) collar on adult Bornean elephant (Elephas Maximus) named *Seri*. Metrics include latitude (lat), longitude (long), temperature (Temperature), external temperature (ExtTemp), activity, speed, direction, covariance (Cov), horizontal dilution of precision (HDOP), distance, and count. The latitude and longitude data show minimal variation, with averages of 5.20° and 118.66° respectively, indicating a specific

geographic area. The mean temperature is 29.20°C, with a wide range from -37.00°C to 60.50°C, suggesting potential anomalies or extreme environmental conditions. Metrics such as external temperature, activity, speed, and direction show uniformity, with all values at 0, possibly indicating static conditions or missing data. Covariance and HDOP have average values of 1.23 and 2.21 respectively, highlighting variable GPS signal quality. Distance values range widely, with an average of 273.89 units, and the Count metric spans from 2199 to 11366, indicating diverse data recording frequencies. The historical elephant *Seri* knowledge graph 4.1 was extracted from FooDS and contains 202,885 triples.

Table 4.1 Summary Statistics of GPS Data and Related Metrics

	lat	long	Temperature	ExtTemp	Activity	Speed	Direction	Cov	HDOP	Distance	Count
count	9168.00	9168.00	9168.00	9168.0	9168.0	9168.0	9168.0	9168.00	9168.00	9168.00	9168.00
mean	5.20	118.66	29.20	0.0	0.0	0.0	0.0	1.23	2.21	273.89	6782.50
std	0.10	0.11	1.93	0.0	0.0	0.0	0.0	1.89	2.07	429.72	2646.72
min	5.01	118.44	-37.00	0.0	0.0	0.0	0.0	0.00	0.00	0.00	2199.00
max	5.38	118.95	60.50	0.0	0.0	0.0	0.0	5.00	21.00	8572.00	11366.00

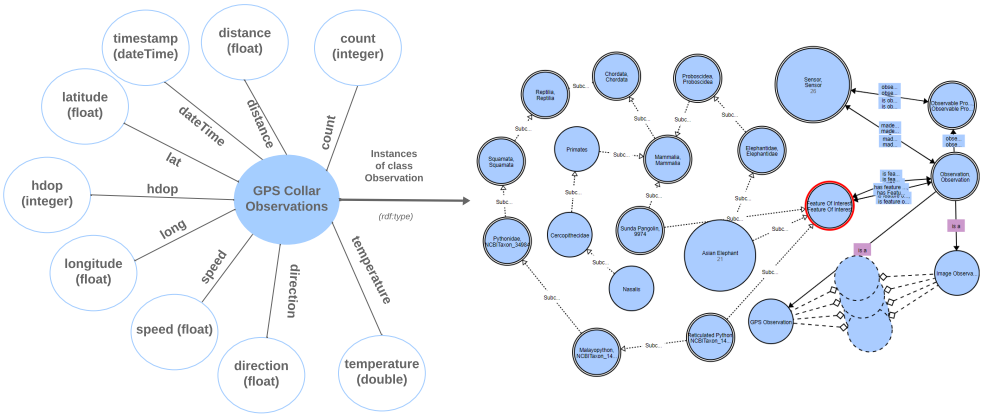


Figure 4.1 Elephant Seri's ontology-based knowledge graph

4.3 Geo-location Prediction with Deep Learning

A neural network model was developed to predict geo-location attributes based on data extracted from the ontology-based knowledge graph. The selected features—localDate, localTime, latitude, and longitude—were chosen for their critical role in capturing both temporal and spatial dimensions of movement. Temporal features such as localDate and localTime provide essential information about when an event or movement occurred, while latitude and longitude define the exact geographic location, making them key predictors for

modelling movement patterns. These features were extracted from the knowledge graph using the rdflib library and SPARQL queries.

4.3.1 Data Extraction and Preprocessing

SPARQL Protocol and RDF Query Language [13] was employed (Algorithm 4.3.1) to retrieve relevant attributes from the RDF graph. The query extracted geo-location data points representing observations for latitude, longitude, date, and time. The extracted data were converted into numerical values for further preprocessing.

Listing 4.1: Query to extract elephant geo-location data including latitude, longitude, date, and time.

```
SELECT *
{?observation   a <https://w3id.org/def/foo#gPSObservation>;
    <http://www.w3.org/2003/01/geo/wgs84_pos#latitude> ?lat;
    <http://www.w3.org/2003/01/geo/wgs84_pos#longitude> ?long;
    <https://w3id.org/def/foo#localDate>   ?localDate;
    <https://w3id.org/def/foo#localTime>   ?localTime. }
```

The retrieved features were then processed to generate feature vectors (‘day‘, ‘month‘, ‘year‘, ‘hour‘) and label vectors (‘latitude‘, ‘longitude‘). The ‘NumPy‘ library was used to convert these vectors into float-compatible arrays, enabling compatibility with the neural network. The dataset was split into training, validation, and testing subsets. Initially, 20% of the data was reserved for testing, while the remaining 80% was split further into 60% training and 20% validation sets.

4.3.2 Architecture and Training

A sequential neural network model was developed using TensorFlow and Keras to predict continuous target variables. The model architecture consisted of an input layer that accepted four features (‘date‘, ‘time‘, ‘longitude‘, and ‘latitude‘), followed by two hidden layers with 128 and 64 neurons, respectively. Each hidden layer used the Rectified Linear Unit (ReLU) activation function to learn non-linear patterns in the data. The output layer employed a linear activation function, enabling precise predictions of the target values (‘date‘, ‘time‘, ‘longitude‘, and ‘latitude‘).

The model was compiled using the Adam optimiser and Root Mean Squared Error (RMSE) as the loss function, quantifying prediction accuracy. The model training was conducted over 500 epochs with a batch size of 32, incorporating validation data to monitor performance and mitigate overfitting. The trained model achieved precise geo-location predictions, evaluated using RMSE. The predictions were transformed into RDF graphs and

integrated into the ontology-based knowledge graph, enriching it for rule-based reasoning. Figure 4.2 illustrates the predictive framework, and pseudocode in 4.3.2 outlines the entire workflow.

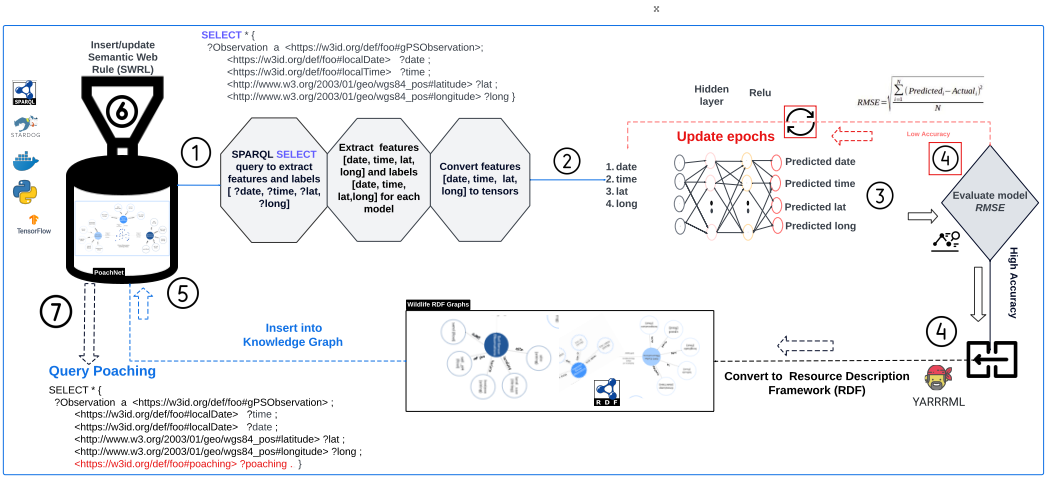


Figure 4.2 POachNet: End-to-end predictive framework featuring RDF data extraction and deep learning and showcasing the integration of the ontology-based knowledge graphs with deep learning. The framework consists of a sequential neural network for predicting elephants' future geo-location. The network comprises an input layer with a shape matching the dataset's four features, followed by two hidden layers employing the Rectified Linear Unit (ReLU) activation function. The output layer uses a linear activation function. The model's performance was assessed using the Root Mean Square Error (RMSE) metric, and accurate predictions were mapped back to their original RDF format.

4.4 Rule-based Reasoning for Poaching Prediction

In rule-based reasoning, each rule comprises conditions and consequents (i.e., outcomes or actions). When the system recognises that the conditions in the antecedents are met, it triggers the action or inference specified in the consequent. This approach is well-suited to scenarios requiring clear and interpretable logic, such as implementing conservation protocols based on animal movements near high-risk areas like plantations or logging zones. Rule-based systems are also adept at handling symbolic representations, making them compatible with semantic reasoning tools like the Semantic Web Rule Language (SWRL), often used in conservation studies to manage and apply rules in ecological knowledge graphs. Using this reasoning framework, specific indicators, such as a binary poaching indicator, can be created to represent the likelihood of poaching occurrences. These indicators can then be incorporated into knowledge graphs, enhancing data-driven decision support systems that

Code Snippet 1: Data Extraction and Regression Model Training

```
Part 1 – Data Extraction ;
Function ParseRDF(graph_file, format):
    graph = LoadGraph(graph_file, format);
    return graph;
Function ExtractFeatures(graph, query):
    features, labels = [], [];
    results = ExecuteQuery(graph, query);
    foreach row in results do
        features.append([row.long, row.lat, row.date, row.time]);
        labels.append([row.long, row.lat, row.date, row.time]);
    return features, labels;
Function SplitData(features, labels, train_size, val_size):
    train_data, test_data = TrainTestSplit(features, labels, train_size);
    train_set, val_set = TrainTestSplit(train_data, val_data, val_size);
    return train_set, val_set, test_data;
graph = ParseRDF("SeriKG.rdf", "ttl");
features, labels = ExtractFeatures(graph, SPARQL_query);
train, val, test = SplitData(features, labels, 0.8, 0.20);
Part 2 – Model Training ;
Function TrainModel(train_data, val_data):
    model = Sequential();
    model.add(Dense(128, activation=ReLU));
    model.add(Dense(64, activation=ReLU));
    model.add(Dense(2, activation=linear));
    model.compile(optimizer=Adam, loss=MeanSquaredError);
    history = model.fit(train_data, validation_data=val_data, epochs=1000);
    return model, history;
model, history = TrainModel(train, val);
Part 3 – Evaluation and Visualisation ;
predictions = model.predict(test_data.features);
RMSE = CalculateRMSE(predictions, test_data.labels);
VisualisePredictions(test_data.labels, predictions);
Part 4 – RDF Graph Enrichment ;
Function MapCSVToRDF(csv_file, ontology, format):
    rdf_graph = LoadOntology(ontology);
    foreach row in csv_file do
        MapRowToTriples(rdf_graph, row);
    SaveGraph(rdf_graph, "SeriKG.rdf", format);
MapCSVToRDF("Seri.csv", "foo.ttl", "ttl");
```

assist conservationists in anticipating and mitigating risks to wildlife populations. Borenean elephants eat diverse vegetation including palm leaves [1] [2] [77, 56]. Monitoring their diet and vegetation choices helps identify danger and poaching hotspots (e.g. oil palm plantations). To predict poaching, Semantic Web Rule Language (SWRL) was used. That is, creating a rule to insert in the ontology-based knowledge graph - inspired by insights from biologists and conservationists at Danau Girang Field Centre (DGFC).

Our experiment created a specific rule (refer to Rule 5) to anticipate poaching activities. This rule predicts poaching based on an elephant's proximity to a designated hazardous area, like an oil palm plantation. The criterion states that if an elephant, equipped with a GPS tracker and termed elephant *Seri*, is near oil palm plantations – areas marked as hazardous owing to previous poaching/poisoning incidents. Thus, there is an increased likelihood that the elephant will be poached. Rule-based semantic reasoning ability led to a new binary poaching indicator in the knowledge graph database (triple store), where '1' indicates potential poaching and '0' indicates its absence.

To determine if an elephant is within a 5 km radius of the oil palm location (see Figure 4.3), I created buffer zones with a radius of 5 km between the oil palm plantation and the elephant geo-location. Figure 4.4 visualises the 5km buffer zones used to formulate the semantic rule.

The haversine formula was applied to calculate the distance between the two points. The elephant can be marked as potentially poaching if the calculated distance is less than or equal to 5 km. The Haversine formula is given by equation 4.4:

$$a = \sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_1)\cdot\cos(\text{lat}_2)\cdot\sin^2\left(\frac{\Delta\text{long}}{2}\right)$$

$$c = 2\cdot\text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$d = R\cdot c$$

Where: Δlat is the difference in latitude , Δlong is the difference in longitude, lat_1 and lat_2 are the latitudes of the two points, R is the radius of the Earth (mean radius = 6,371 km), d is the distance between the two points.

[1] borneomammals.online/2018/10/15/bornean-elephant-feeding-in-oil-palm-at-tabin/

[2] rainforest-rescue.org/petitions/905/malaysia-pygmy-elephants-poisoned-for-palm-oil

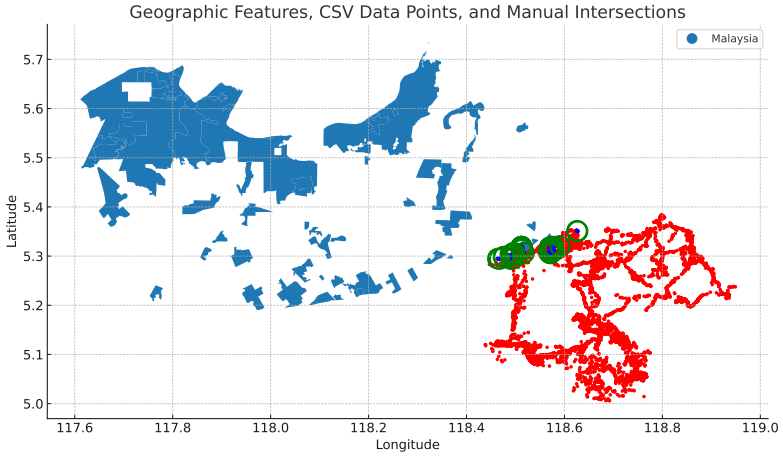


Figure 4.3 The map shows the geo-points (in green) intersection between the oil palm plantation (in blue) and the elephant movements (in red)

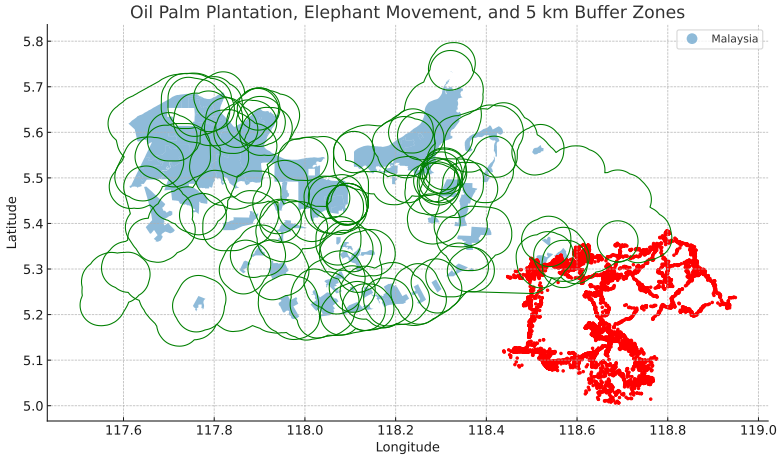


Figure 4.4 The figure displays a 5 km buffer zone (in green) around each geographic feature of an oil palm plantation. These zones are overlaid with the original geographic features (shown with slight transparency) and the points from the elephant movements (in red).

Listing 4.2: Detect Poaching Observations Near Oil Palm Plantations Rule

```
INSERT {
  ?s    a <https://w3id.org/def/foo#gPSObservation>;
          <https://w3id.org/def/foo#poaching> ?poaching.
}
WHERE {
?s   a   <https://w3id.org/def/foo#gPSObservation>;
         <http://www.w3.org/2003/01/geo/wgs84_pos#latitude> ?lat;
         <http://www.w3.org/2003/01/geo/wgs84_pos#longitude> ?long.

  # Retrieve plantation details
  <https://w3id.org/def/foo#plantation> a <https://w3id.org/def/foo#OilPalmPlantation>;
                                          <http://www.w3.org/2003/01/geo/wgs84_pos#latitude> ?plantationLat;
                                          <http://www.w3.org/2003/01/geo/wgs84_pos#longitude> ?plantationLong.

  # Convert coordinates to float (if stored as literals)
  BIND(xsd:float(?lat) AS ?latitude)
  BIND(xsd:float(?long) AS ?longitude)
  BIND(xsd:float(?plantationLat) AS ?oilpalmLat)
  BIND(xsd:float(?plantationLong) AS ?oilpalmLong)

  # Calculate distance using the Haversine formula
  BIND(6371 * 2 * ASIN(SQRT(
    POW(SIN((?latitude - ?oilpalmLat) * PI() / 180 / 2), 2) +
    COS(?oilpalmLat * PI() / 180) * COS(?latitude * PI() / 180) *
    POW(SIN((?longitude - ?oilpalmLong) * PI() / 180 / 2), 2)
  )) AS ?distance)

  # Determine poaching based on the calculated distance
  BIND(IF(?distance <= 5, 1, 0) AS ?poaching. }
```

Listing 4.3: Query to retrieve poaching status in a format of turtle graph with the geo-location coordinates, local data and poaching likelihood.

```
CONSTRUCT WHERE {
  ?Observation   a   <https://w3id.org/def/foo#gPSObservation>;
          <https://w3id.org/def/foo#localDate>    ?LocalDate ;
          <http://www.w3.org/2003/01/geo/wgs84_pos#latitude> ?lat ;
          <http://www.w3.org/2003/01/geo/wgs84_pos#longitude> ?long ;
          <https://w3id.org/def/foo#poaching> ?poaching.   }
```

4.5 Results

This section presents the geo-location prediction results. The foundation of this approach is FooDS, now publicly accessible online at (w3id.org/def/fooDS). However, elephant *Seri* GPS Observations dataset used in this research is kept confidential due to its sensitive nature. The graph injected with Semantic Web Rule Language (SWRL) enabled the semantic reasoning about poaching and introduced new triples to assert poaching likelihood. The query results fed into the deep learning models demonstrated high accuracy and compatibility with machine learning formats. To evaluate the accuracy of the geo-location predictions, the Root Mean Square Error (RMSE) was used. Deep learning model was trained using Tensorflow within a Docker environment. The remote computer machine hosting the model

had an Intel Xeon W-2245 processor, an NVIDIA GPU RTX A6000 with 48GB memory, and 128GB of DDR4 RAM.

4.5.1 Geo-locations Prediction Result

The proposed neural network model for the geo-location prediction is a linear model and built using the TensorFlow and Keras frameworks. The model contained an input layer intended to accommodate four critical features (data, time, longitude and latitude) related to the geographical positioning of elephant *Seri*. The network also includes two subsequent dense layers, containing 128 and 64 neurons, respectively, using the Rectified Linear Unit (ReLU) activation function to capture nonlinear patterns in the data effectively. In other words, the model is an output layer with two neurons, employing a linear activation function. Such configuration is well-suited for regression tasks of continuous outputs (i.e., longitude and latitude). The data used contained 9168 observations, and their distribution is shown in Figure 4.2 step 3. This model underwent multiple training epochs with a batch size of 32. It achieved its highest accuracy at 500 epochs, registering an average geospatial RMSE of 0.0166 for elephant *Seri* GPS observations dataset.

4.5.2 Evaluation

To evaluate predictive methods on *Seri* GPS collar data, I used its dataset in CSV format containing (date, time, longitude, latitude) features. The goal was to predict spatial-temporal coordinates (date, time, longitude, latitude) and compare the performance of three models: linear regression, polynomial regression, and vector autoregression (VAR). The performance was assessed using the average root mean square error (RMSE).

4.5.3 Data Preprocessing

The *independent variables* in this analysis are the input features used to make predictions. These include the day, month, and year, all of which are extracted from the 'LocalDate'. These temporal features provide the contextual information necessary for the models to make accurate predictions.

The *dependent variables*, or targets, are the outputs the models aim to predict. These include geospatial coordinates such as latitude ('lat') and longitude ('long'), as well as temporal features like the day, represented as a numeric value indicating the day of the month, and time, converted into a numeric representation of seconds since midnight (e.g.,

‘12:34:56‘ becomes ‘45296‘ seconds). Together, these outputs capture both spatial and temporal dynamics.

The data were divided into training and testing sets to ensure fair evaluation, and models were trained and tested sequentially to respect the temporal structure of the data.

4.5.4 Models and Results

1. Linear Regression: Applied simple linear regression using day as the predictor for both lat and long. Evaluated using 5-fold time-*Seri*es cross-validation.

2. Polynomial Regression:Incorporated polynomial features (degree 4) to account for non-linear relationships. Similarly validated with 5-fold time-*Seri*es cross-validation.

3. Vector Autoregression (VAR): Used both latitude and longitude as a multivariate time *Seri*es for temporal forecasting. Reserved a portion of the dataset for out-of-sample prediction and RMSE calculation.

The RMSE results for all models are summarised in Table 4.5.4. From the negatively-oriented RMSE scores (where a lower score indicates better performance), the linear regression model demonstrated strong performance for predicting latitude and longitude, with RMSE values of 0.123 and 0.164, respectively. Notably, the linear regression model also performed exceptionally well for predicting the day, achieving an RMSE close to zero, but struggled with time predictions, yielding an RMSE of approximately 25,186 seconds.

The polynomial regression model, while offering a more complex representation, exhibited higher RMSE values for latitude (2.396) and longitude (1.050) predictions. It achieved a near-perfect prediction for the day (RMSE: 7.2e-06) but produced the highest error for time predictions, with an RMSE of 483,988 seconds. This suggests that the added complexity of the polynomial model may have led to overfitting or inefficiency for these particular features.

In comparison, the VAR model excelled in predicting longitude, with the lowest RMSE of 0.089. It also performed reasonably well for latitude predictions (RMSE: 0.222) but struggled with day and time predictions, with RMSE values of 8.69 and 25,093 seconds, respectively. These results indicate that the VAR model effectively captures temporal dependencies for geospatial coordinates but may require additional feature refinement for accurate temporal predictions.

However, the neural network built with TensorFlow and Keras trained on the same *Seri* data but in an ontology-based knowledge graph, outperformed all other models. The neural network model achieved test RMSE values of 0.0247 for longitude, 0.0084 for latitude, 0.0123 for ‘localDate‘, and 0.0086 for ‘localTime‘. These results demonstrate the effectiveness of

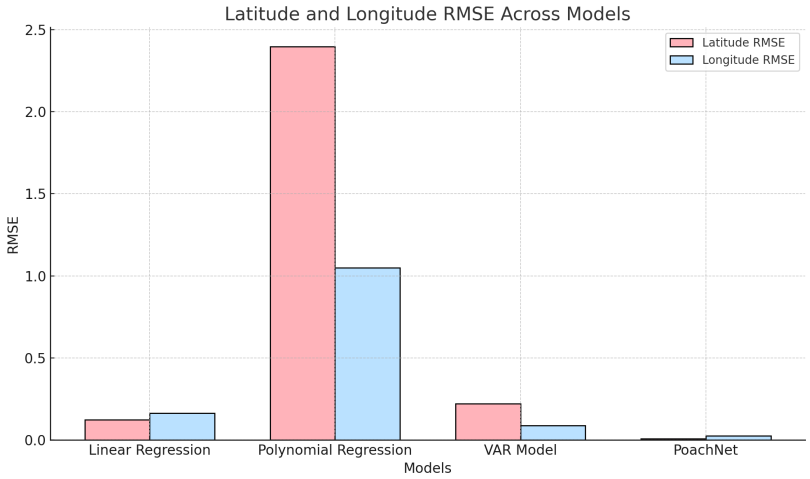


Figure 4.5 The chart visually compares the performance of various predictive models, specifically focusing on their Root Mean Square Error (RMSE) values, a standard measure of prediction accuracy. The models, including 'Linear Regression', 'Polynomial Regression', 'VAR', and 'PoachNet'.

leveraging a knowledge graph representation and deep learning methods for highly accurate geospatial and temporal predictions. Codes are available in Github.

Table 4.2 Comparison of RMSE between PoachNet and State-of-the-Art Models

Model	**Latitude RMSE**	**Longitude RMSE**	**Average RMSE**
Linear Regression	0.123	0.164	0.144
Polynomial Regression	2.396	1.050	1.723
VAR Model	0.222	0.089	0.156
PoachNet	0.0084	0.0247	0.0166

4.6 Discussion

The loss of forest elephants and their dispersal from poaching or habitat loss and fragmentation [106] could lead to reduced forest diversity, the inability of elephants to colonise new or deforested areas, and potentially reduced carbon stocks. Combating poaching in Sabah is a priority, and various organisations, including Sabah Wildlife Department and Sabah Forestry Department with the support of Danau Girang Field Centre and WWF-Malaysia, are working to protect the Bornean elephant and many other species. The Bornean Elephant Action Plan for Sabah 2020-2029 is a ten-year plan approved by the state government of Sabah to conserve the Bornean elephant population and many other species. The plan has

four main objectives: improve protection and reduce elephant deaths, improve landscape connectivity and permeability, ensure the best ex-situ practices for elephant management and conservation, monitor, and predict elephant population trends.

This research differentiates itself from existing views by integrating heterogeneous wildlife data with deep learning on an ontology-based knowledge graph. while prior approaches have primarily focused on specific aspects, such as social network analysis, multimedia data mining, or hierarchical models on ranger patrol data, this methodology offers an interconnected understanding of wildlife dynamics. The results highlight that while linear regression is well-suited for simple relationships in this dataset, and the VAR model shows promise for geospatial predictions, PoachNet surpasses them significantly, showcasing the potential of neural networks combined with knowledge graph techniques. Polynomial regression, despite its theoretical flexibility, did not outperform the simpler models and may require better feature engineering to improve its effectiveness.

PoachNet predictions can assist the strategic resource allocation for anti-poaching efforts. It can also guide the decision to deploy ground truth sensors and motion-activated camera traps in areas most likely to have anticipated poaching crimes.

Research challenges include semantic heterogeneity among diverse data sources, which risks the consistent representation of information in the knowledge graph. Scalability issues may emerge as the knowledge graph expands, necessitating careful resource management. To address scalability issues in the knowledge graphs, several strategies can be recommended. Partitioning the graph into manageable subgraphs and using distributed triple-store databases like Stardog, Neo4j or Amazon Neptune can enhance processing efficiency. Incremental updates minimises reprocessing, while graph compression and summarisation reduce storage demands. Scalable cloud-based storage, optimised query processing with indexing, and the use of high-performance graph algorithms further improve performance. Edge computing can preprocess data near collection points, reducing bandwidth and latency.

Optimising the deep learning algorithms in PoachNet to enhance predictive performance while minimising computational costs can be achieved through model compression techniques such as pruning and quantisation [198]. These approaches reduce the size of deep neural networks while maintaining accuracy, enabling faster inference, reduced storage requirements, and lower training costs. Techniques like low-rank decomposition, knowledge distillation, and lightweight model design can further streamline model deployment, making them more efficient for use in resource-constrained environments [156].

PoachNet can be expanded by integrating additional wildlife data sources such as acoustic sensors, satellite imagery, and crime intelligence. Acoustic sensors can detect gunshots, elephant vocalisations, or vehicle noises associated with poaching crimes. Satellite imagery

can monitor changes in habitat, detect unauthorised human activity, and assess landscape connectivity. Crime intelligence data can add historical context, identifying patterns in poaching incidents and aiding in predicting future hotspots.

4.7 Summary

This chapter introduced PoachNet, a novel tool integrating Semantic Web technologies and deep learning to predict poaching crime and wildlife dynamics. By combining diverse wildlife data into an ontology-based knowledge graph enriched with rule-based reasoning, PoachNet provided a dynamic, hybrid predictive solution for conservation. Custom-built dataset and advanced neural network models accurately predicted elephant geo-locations and potential poaching incidents, achieving an average geospatial RMSE of 0.0166 , surpassing state-of-the-art methods. This approach predicts future elephant geo-locations and uses this information to infer poaching risks based on proximity to identified hazardous areas. PoachNet equips biologists and conservationists with advanced tools for spatiotemporal poaching predictions, offering a transformative paradigm for wildlife crime prevention. While challenges such as semantic heterogeneity, data sensitivity, and ecosystem dynamics persist, the public release of the ontology-based knowledge graph and source code demonstrates the commitment to transparency and collaboration, encouraging the research community to collaborate with us and build upon this work.

Chapter 5

Extending FooDS's Semantic Web Framework to Data Marketplaces

This chapter addresses the fourth research question (RQ4): *Can the Linked Data Store's semantic web data management approach be generalised to another domain for various purposes?*

This chapter examines whether FooDS's semantic web data management approach can be applied to other fields, with a focus on IoT data marketplaces. These marketplaces represent a new concept designed to meet consumer data needs by encouraging data sharing and lowering acquisition costs. However, existing marketplaces, such as SynchroniCity, lack options for selective data purchasing, requiring consumers to buy entire datasets, which can be costly and inefficient. FooDS's approach was adapted to an IoT data marketplace, enabling selective querying of annotated IoT data. This adaptation allowed users to access only the data they needed, reducing costs and improving efficiency. An ontology, developed with experts input, reused existing ontology and was populated with six different sensor datasets to create ontology-based knowledge graphs. Semantic Web Rule Language (SWRL) reasoning was applied to three use cases, demonstrating its ability to efficiently manage rules, store data at the edge, and provide remote access through SPARQL queries without straining resources. This case study illustrates the potential of FooDS's approach to be generalised and applied in other domains.

5.1 Introduction

Over the last decade, many cities have initiated projects that deploy different sensors for various reasons. One popular application domain is environmental monitoring. After

accomplishing the primary objectives, such data are often discarded or stored somewhere where access can be difficult outside the initial project. There often needs to be a mechanism (or motivation) to share data with outside parties (other than the organisation that deploys the sensing infrastructure). This approach leads to a waste of resources and can limit the potential benefits derived from such data. Internet of Things (IoT) data marketplaces for smart cities are being proposed as a solution to address this challenge. *Urban data Exchange*[1] is one such data marketplace aimed at facilitating businesses to develop IoT and AI-enabled services to improve citizens' lives and grow local economies. Data marketplaces have received limited attention in the academic community. However, the buying and selling of data have taken place for a long time, especially within the business-to-business (B2B) context. Initially, these data transactions took place offline between companies and their alliances. Data have been widely sold across various domains, such as travel, advertising, and insurance. This chapter addresses the fourth research question (RQ4): *Can the Linked Data Observatory 's approach be generalised to another domain for various purposes?* Several key contributions are presented and linked to the research question:

- An ontology, aggregating well-known ontologies to model sensor data in IoT marketplaces, is proposed. Design decisions made during the ontology engineering process are discussed, highlighting trade-offs and identifying good practices observed in existing efforts.
- A unique technique for on-demand data offer creation is introduced. Custom data requests (i.e., data orders) are enabled, allowing buyers to consider four aspects: location, data type, date/time, and service level agreement.
- The utility of knowledge engineering, including reasoning and inferencing, is demonstrated through a series of use cases in the context of data marketplaces. Different levels of knowledge engineering approaches—dataset level, market level, and buyer level—are presented, along with their utility and associated costs, such as computational complexity.
- The performance of the proposed approach is evaluated across three different data marketplace setups. Various parameters are measured, and several recommendations for future marketplace deployments are extracted.

[1] https://urbandata.exchange/

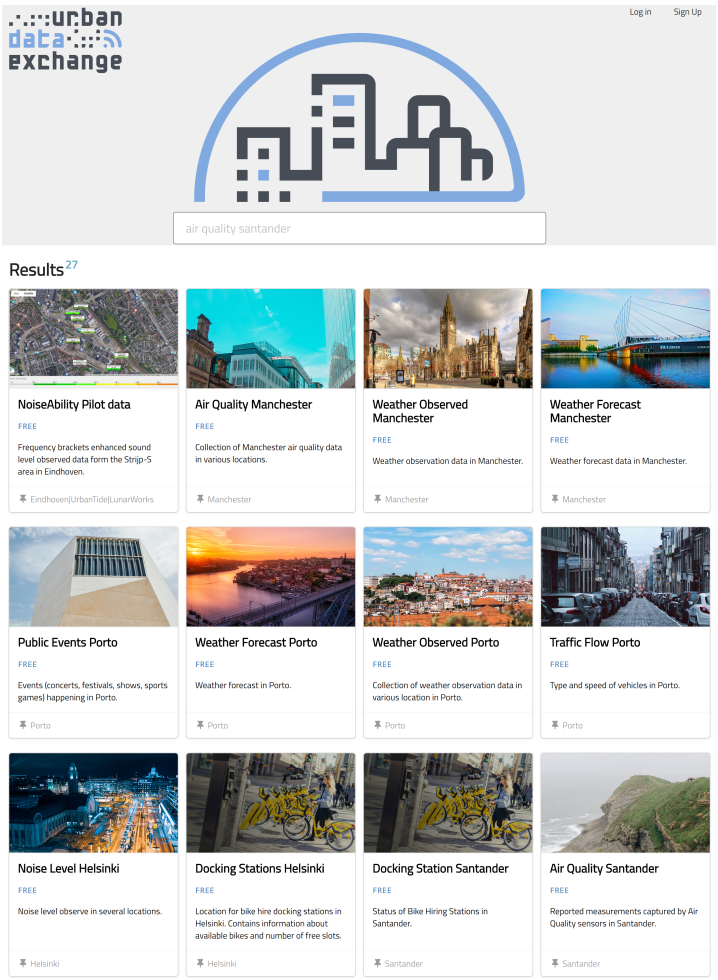


Figure 5.1 Urban Data Marketplace (Current Approach): FIWARE data models are used to organise the data into datasets. They have been harmonised to enable data portability for different applications, including Smart Cities, Smart Agrifood, Smart Environment, Smart Energy, Smart Water, and others. The key weakness of this approach is that data buyers need to buy the entire dataset (e.g., Noise Level Helsinki) whether they need the entire dataset or not. This approach leads to higher data prices.

5.2 Motivation and the Problem Definition

Currently, IoT data marketplaces sell data per entire dataset, as shown in Figure 5.1. For example, potential buyers could buy weather forecast data and parking status data in bulk. Each of the datasets may contain multiple pieces of data packed together in a pre-defined manner (e.g., `temperature`, `relativeHumidity` may be included in the Weather Forecast data offer). There are multiple problems with this approach.

Problem 1: Higher network communication and bandwidth requirement: ($\uparrow D_i^p \propto D_i^b \uparrow$) In the current approach, data offers are sold as pre-defined data bundles. There is no way to limit the number of data within a bundle that a buyer acquires (whether the buyer may request past (archival) data or future data (as a subscription)). The consequence of this approach is higher network download time and cost. For example, if a data consumer wants bicycle docking station data in Santander, they will need to buy the entire `Docking Stations - SAN` dataset whether they need the entire dataset or not. Therefore, there is a positive correlation between network communication, bandwidth requirement and volume of data. Given that the price of a data offer i is D_i^p and the bandwidth required is D_i^b, the price is proportional to bandwidth.

Problem 2: Difficulties in pricing which leads to higher prices: ($\uparrow D_i^p \propto D_i^v \uparrow$) Currently, each data bundle comprises large volumes of data. The cost of acquisition for a large amount of data is high. Therefore, the cost of the bundle has to be high as well. In a data marketplace, the price of a data offer has to cover the cost of data acquisition plus a profit margin. Therefore, there is a positive correlation between data prices and volume. Given that the price of a data offer i is D_i^p and volume is D_i^v, price is proportional to volume.

Problem 3: Information overload for data consumers: ($\uparrow D_i^v \propto D_i^{pp} \uparrow$) .
In data science projects, 80% of time and effort is often devoted towards preparing the data (i.e., acquiring, cleaning, transforming, etc.), and only 20% is used to do the actual analysis. Therefore, the larger the buyers' dataset, the more effort they need to prepare and filter the relevant data. Assume that a given data analysis task is related to `weekends` data (e.g., parking slot status). First, data scientists need to query the entire dataset, remove the data related to `weekdays` and select only the data related to `weekends`. Therefore, the current bulk pre-defined data offering approach unnecessarily increases data scientists' workload (i.e., data consumers'). Given the size of a data offer, i is D_i^v, the cost of data pre-processing is D_i^{pp} – the cost of data pre-processing is proportional to volume.

Problem 4: Limited data discovery capabilities: Currently, IoT data marketplaces organise datasets by type (broadly) and location. For example, it is usually up to the data seller to bundle the data into an offering as they see fit, as shown in Figure 1. Here, data offers are pre-defined and static without any mechanism to request customised data. Data search primarily relies on location. There is no way for data consumers to acquire `traffic data` in `London` on `rainy` days over the `last three years` in the current data marketplace scenario.

5.2.1 Design Principles and Architecture

Design principles were suggested to help data consumers communicate the data they need. These efforts yielded Competency Questions (CQs) for our data model.

Design Principles

Twenty-one participants were enlisted with backgrounds in computer and data science to extract the following design principles. The identified expressed their priorities using question terms: (1) where, (2) what, (3) when, and (4) how. Let us explain each of these design principles with a concrete example.

Selected participants were considered for their expertise in three key areas: (i) their understanding of the IoT domain, (ii) their knowledge of semantic web technology for the IoT from an end-user perspective, and (iii) their proficiency in semantic data modelling. I reviewed their qualifications, expertise, and past experiences, sourcing them from professional networks, academic institutions, conferences, and events. The participants' questionnaire responses supplemented us with the necessary data for these filters to make deductions. For instance, one of the questions asked, "*Do you believe that bikes available for hire in London city will be less accessible than usual on a sunny day?"* The response showed that over 70% of the participants agreed with this statement. Consequently, logical rules were established to infer sunny days based on the decreased availability of bikes for hire. Moreover, the participants' responses revealed interesting insights about their demographics. Most participants have achieved a minimum of a bachelor's degree, with approximately 30% of them having pursued further education at the postgraduate level. The majority of participants, as such, have direct experience in applying semantic web technology to handle and analyze IoT data (e.g., database management, AI and machine learning). The distribution of years of experience in the current field is quite diverse, with 25% having 1-2 years, 30% having 3-5 years, and a sizeable proportion having more than ten years. Participants also have a wide range of ages, with the majority in the 25-34 and 35-44 age groups, respectively. Figure 5.2 and 5.3 show the study participants' educational background and demographic information.

Architecture

The IoT data marketplaces need to be distributed in nature. Data owners are expected to store and manage data items and only share their metadata with brokers such as the IoT data marketplaces. Consequently, when a broker receives a request, it knows from where to gather the data (or to decide whether it is possible or not to fulfill the request). Figure 5.4 depicts the architecture of the IoT data marketplace, and details are presented

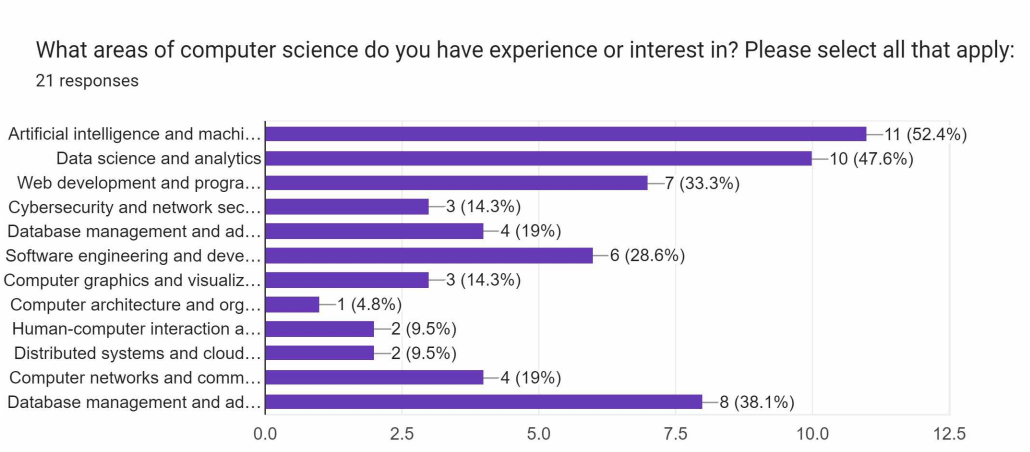


Figure 5.2 Participants education and experience background in Computer Science.

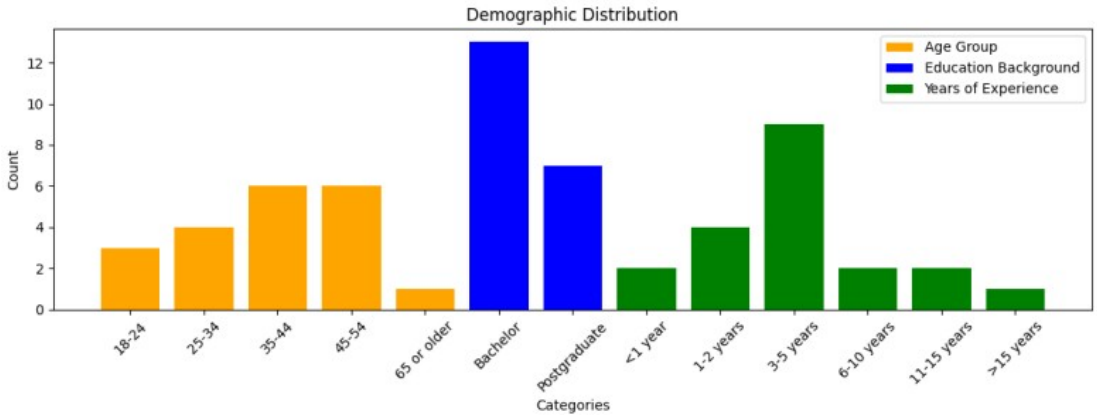


Figure 5.3 Participants demographic information.

here (https://gitlab.com/synchronicity-iot). In summary, we have developed a user interface that allows data consumers to build their data requests. Each data request was organised using a standardised JSON schema and send it to the validation engine. The validation engine determines whether the IoT data marketplace can fulfill a given request based on the available metadata. Then, one or multiple SPARQL queries will be generated based on the requirements of the data request (and depending on where the actual data reside).

5.3 Data Marketplace Design

To address the requirements stated in 5.2.1, an IoT data marketplace was proposed that allows potential data consumers to buy *only the data points/records they need to solve a given problem*. Pre-defining many data offerings is not feasible; therefore, the best approach is to allow consumers to create their data offerings (i.e., data requests). To illustrate the notion, assume a new tourism company is interested in purchasing various data entities to build an

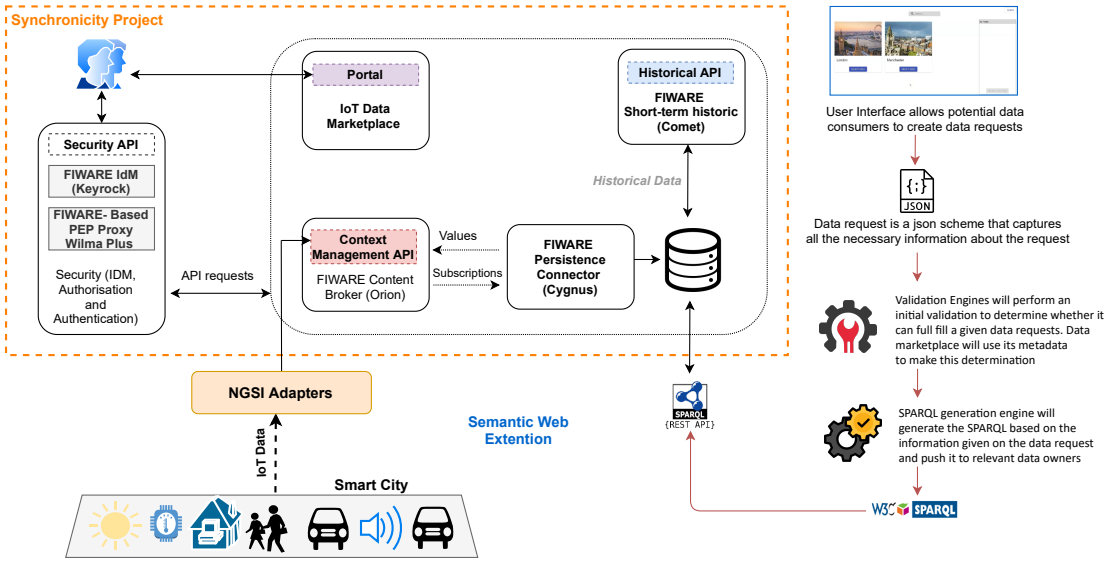


Figure 5.4 Semantically Enhanced IoT Data Marketplace Architecture

AI decision support system. These entities may contain specific observations about local attractions such as beaches, museums, parking spots, and bike docking stations. Thus, the organisation prepares a single order that includes the data records from the datasets instead of acquiring the entire dataset for each entity separately. Nonetheless, this method has several drawbacks in terms of the following: (see Table 5.1).

- Data pricing could be complicated because each data source may price its observations-depending on their size and novelty. In addition, the broker fees and any variable extra charges have to be carefully calculated and added to the total bill.
- **Data publishing** could raise privacy and ownership concerns because data providers may have different privacy policies and credit preferences. Therefore, appropriately tailored privacy and data ownership agreements must exist to satisfy all parties.

Table 5.1 Possible Scenarios

Criteria	Current Pre-Defined Data Offering Approach	Proposed On-Demand Data Offering Approach
• Pricing Structure	Simple	Could be complicated
• Pricing Fairness	Less fair	More fair
• Data Discovery	Less discoverable	More discoverable
• Publishing Complexity	Easy and simple to publish	Could be complicated to publish
• Data Preparation Complexity (from a data consumer perspective)	Higher (as large datasets need to be processed and filtered)	Less (as data is already processed and filtered)

5.3.1 Data Model

Proposed data model comprised a foundational ontology instantiated with six heterogeneous datasets. The ontology aims to describe sensor data in the IoT data marketplace. Following the NeOn methodology [103, 102, 239], the ontology was built. Although there are numerous other ontology development methodologies [210, 87, 242, 190, 211], NeOn was selected as it has multiple modular scenarios to choose from and adapt to our current requirement. Figure 5.5 shows the ontology development's life cycle. I adopted NeOn's first, second and third scenarios. The first scenario outputs the Ontology Requirement Specification Document (ORSD). Then, I identified the non-functional requirements from the second scenario based on the ORSD. I , then followed the process of reusing existing ontological resources from the third scenario. Following that, the ontology was implemented and evaluated using various tools. Figure 5.6 depicts the proposed core ontology, Table 5.2 discusses the key characteristics of this ontology, and Figure 5.7 shows the ontology instantiated with six sensor datasets.

5.3.2 Ontology Requirements

The first step in developing the ontology was to gather the required information. Here, the information collected at the design stage (see section 5.2.1 were coded into Competency Questions (CQs) to develop the ontology. During this step, the ORSD was produced, which contains the conceptual building blocks for the ontology as follows:

- Ontology purpose: To describe the sensor data.

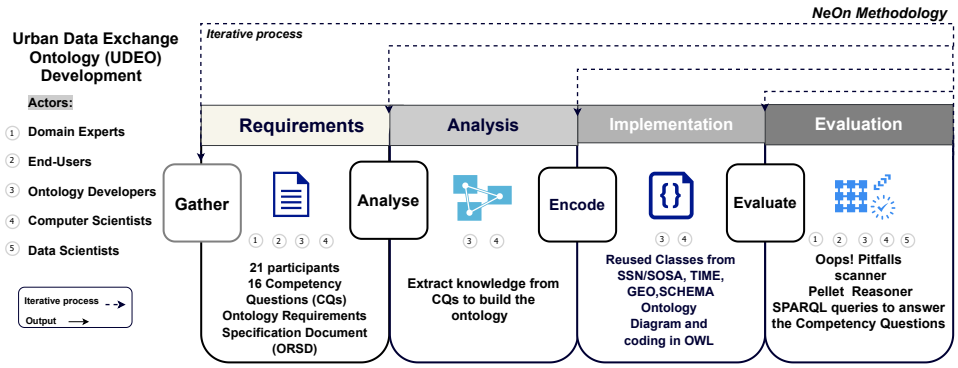


Figure 5.5 Neon Methodology for developing the proposed Urban Data Exchange Ontology

- Ontology scope: The Internet of Things (IoT).
- Ontology implementation language: OWL2 Web Ontology Language.
- Ontology intended users: Small to medium businesses (SMEs).
- Ontology non-functional requirements: To List elements that must be included in the ontology, such as IoT geospatial and time classes.
- Ontology functional requirements: To contain the Competency Questions (CQs) that build and validate the ontology.

5.3.3 Ontology Analysis

In this step, the Competency Questions (CQs) were revised from the requirements phase and extracted knowledge to implement the ontology. Reusing classes from mature ontologies to develop a new ontology that models the following concepts was decided : 1) sensor data observations, 2) sensing infrastructure, 3) location, 4) temporal aspects, and 5) units of data.

To find suitable state-of-the-art ontologies, I searched Google Scholar [105] and the BioPortal repository [1], as well as other scholarly websites and ontology repositories, with inclusion criteria that the publication date had to be between 2015 and 2020. Different search terms were used to perform the search: "sensor data ontology", "semantic modelling for sensor data", "semantic IoT data", and "IoT ontology" . The outcome of the search yielded six ontologies that are commonly used to model sensor data, such as the Semantic Sensor Network Ontology (SSN) [256].

Among the shortlisted ontologies is the SAREF [61] ontology that describes smart appliances and related IoT devices and services, which may not be the most suitable ontology

for modelling sensor data observation. Additionally, the IoT-Lite ontology [24] provides a basic set of classes and properties for describing IoT devices, sensors, and actuators. However, more than this may be needed for the ontology use cases, as classes are required to model the sensor's observation and properties rather than the sensor alone. The W3C Web of Things (WoT) ontology[2] is a flexible, modular ontology that can be changed to fit different use cases and allow different IoT systems and domains to work together. WoT focuses on different aspects of IoT devices and services. However, its flexibility and generality can also make it challenging to adapt to our requirements. For example, one of the WoT classes, "Thing", models the IoT device, the service, or the data source. In contrast, the class "Sensor" is more suitable for modelling the sensor's observation.

The FIESTA-IoT ontology [3], as such, models IoT-related concepts but has more entities than needed. It borrows classes from the SSN ontology (Version 1) [52], the W3C Web of Things (WoT) Thing Description, and the oneM2M standard[4]. The IoT-Semantics Ontology is another flexible ontology that lacks sufficient documentation, making it challenging for developers to adapt. After scouring and comparing the state-of-the-art ontologies, reusing concepts from the SSN ontology (Version 2) [256] was decided, mainly for its modular property. SSN ontology (Version 2) integrates three distinct ontologies. That is the SSN ontology (Version 1), the Sensor, Observation, Sample, and Actuator (SOSA) ontology [132], and the Quantities, Units, Dimensions, and Types (QUDT) ontology [192], qualifying it to be the optimal choice for our use case. Further, the new ontology was named the Urban Data Exchange Ontology (UDEO).

The sensor data observations were modelled using the SOSA ontology, which was extracted from SSN (Version 2). The SOSA ontology is designed to be interoperable with other Semantic Web standards like RDF, OWL, and SPARQL, making it easy to integrate with other data sources and applications. Additionally, the SOSA ontology is continuously maintained and updated by a community of researchers and developers, ensuring its relevance to current IoT applications. Its modular and extensible nature allows for easy customisation to suit specific use cases. SOSA also provides sufficient facilities to model sensing infrastructure. For location, this work focuses on outdoor locations that are easily identified using GPS coordinates. Temporal aspects were addressed by storing the timestamp of each observation using the XML DateTime data type (xsd:dateTime). Whilst the incorporation of OWL-Time into UDEO was considered, it was concluded that SPARQL's Time function provides most of OWL-Time's capabilities. Consequently, OWL-Time was omitted from the ontology. For units of data, concepts from the Quantities, Units, Dimensions, and Types Ontology (QUDT)

[2]https://www.w3.org/TR/wot-thing-description11/

[3]http://iot.ee.surrey.ac.uk/ontology/fiesta-iot.owl

[4]https://www.onem2m.org/

were adopted. QUDT is part of the SSN ontology (Version 2) and provides a modular approach to ontology development, ensuring compatibility and interoperability with other systems using QUDT or similar ontologies.

5.3.4 Ontology Implementation

As mentioned in the analysis step, most of the UDEO classes were adopted from the SOSA ontology as shown in Figure5.6. The conceptual ontology model, or the lightweight version, was designed as a UML diagram in draw.io software[5]. The ontology diagram in 5.7 was discussed between UDEO stakeholders before its digitisation. Each of the concepts and relationships modelled in UDEO is listed in Table 5.2. The ontology's digital version was coded in the Protege ontology editor and exported it to a dedicated knowledge graph platform [2].

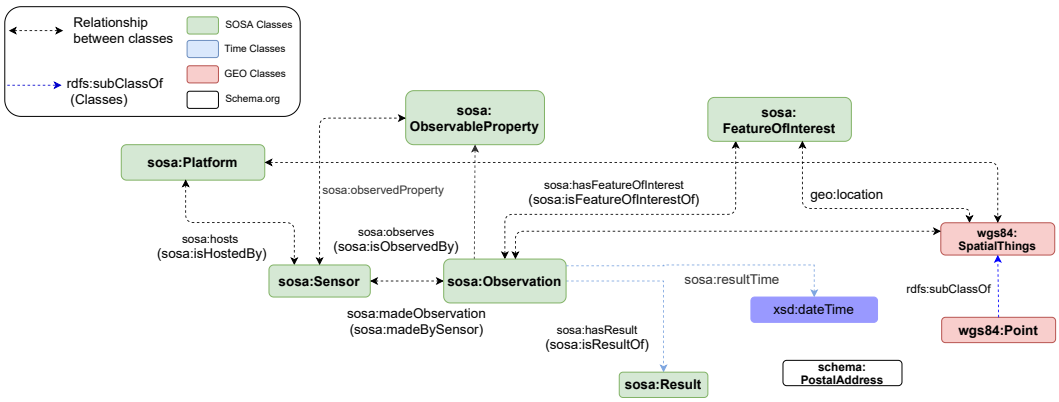


Figure 5.6 The adopted classes for the proposed Urban Data Exchange Ontology (UDEO)

Table 5.2 Proposed data model: Concepts and relationships. [Official definitions are in *Italics*]

Concept	Description
Observation (OWL Class) [sosa]	*Act of carrying out an (Observation) Procedure to estimate or calculate a value of a property of a FeatureOfInterest (e.g., Room).* Observation can be seen as a placeholder that links relevant information together. As illustrated in Figure 5.7, observation can be considered an ID for each data record in our data model. Each raw depicts a data record.
ObservableProperty (OWL Class) [sosa]	*An observable quality (property, characteristic) of a FeatureOfInterest.* (e.g., temperature, humidity, presence)

[5]https://app.diagrams.net/

Sensor (OWL Class) [sosa]	*Device, agent (including humans), or software (simulation) involved in, or implementing, a Procedure.* (e.g., Temperature sensor, humidity sensor, motion sensor). In our model, a unique ID was created for each sensor based on its hosted platform.
Platform (OWL Class) [sosa]	*A Platform is an entity that hosts other entities, particularly Sensors, Actuators, Samplers, and other Platforms.* In UDEO, sensors are attached to different types of platforms, as shown in Figure5.7. I do not necessarily keep track of the exact location of the platform. However, location can be approximately identified by using the feature of interest.
FeatureOfInterest (OWL Class) [sosa]	*The thing whose property is being estimated or calculated in the course of an Observation to arrive at a result, or whose property is being manipulated by an Actuator, or which is being sampled or transformed in the act of Sampling.* In the context of UDEO, *BuidlingSpaces* are the *FeatureOfInterest* (e.g., offices, zones, floors). Most of the sensors are used to observe a property (phenomenon) of a location (e.g., the temperature in a room).
Result (OWL Class) [sosa]	*The Result of an Observation, Actuation, or act of Sampling. To store an observation's simple result value, one can use the hasSimpleResult property.* Result is a placeholder to link related information, such as values and units. The UDEO model stores the data value and its unit type.
resultTime (Datatype Property) [sosa]	*The result time is the instant of time when the Observation, Actuation or Sampling activity was completed.* Each data record in the UDEO system comes with a time stamp.
SpatialThing (OWL Class) [WGS84]	*A class for representing anything with a spatial extent, i.e., size, shape or position.*

5.3.5 Ontology Evaluation

This step evaluates the ontology quality in terms of *structure, semantic representation, and interoperability*. To evaluate the structure and semantic representation, the open-source online scanner, Oops! [211] and Pellet reasoner [231] built-in to Protege were used. Following that, SPARQL queries were executed against the knowledge graph (i.e., UDEO instantiated with IoT datasets) to answer the Competency Questions (CQs).

5.3.6 Experimentation Plan

Our experimentation plan consists of three layers, as shown in Figure 5.18, (i) data sources, (ii) adaptor layer; and (iii) evaluation. The aim is to simulate a data marketplace using the

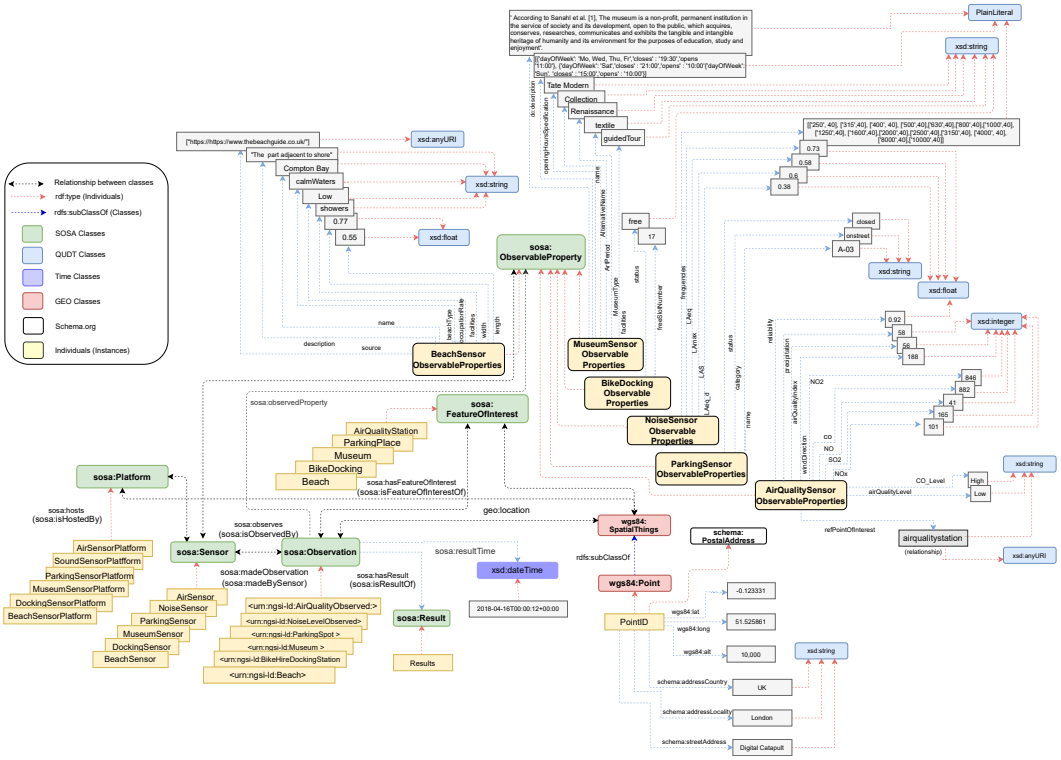


Figure 5.7 The proposed Urban Data Exchange Ontology (UDEO) instantiated with six different sensors' datasets

most practical solution that fits the purpose.

- Data Sources: UDEO was expanded to accommodate six data sources: docking stations, air quality, noise level, parking status, museums, and beaches. Further, synthetic datasets were generated for each data source that adhered to the FIWARE data model structure.
- Adaptor Layer: six datasets were mapped into the Resource Description Framework (RDF) graph - referencing UDEO.
- Evaluation Layer: the data were stored from the adaptive layer in triple-store databases under three different architectures and evaluated each one's performance.

SynchroniCity IoT Data Marketplace receives data as JSON files (modeled using FIWARE standards) from data owners/publishers. SynchroniCity IoT Data Marketplace currently has a limited amount of data. It was not sufficient for us to conduct an experiment to uncover the utility of Semantic Web technologies and their impact on computing infrastructure. Therefore, an algorithm was developed to produce the required number of synthetic

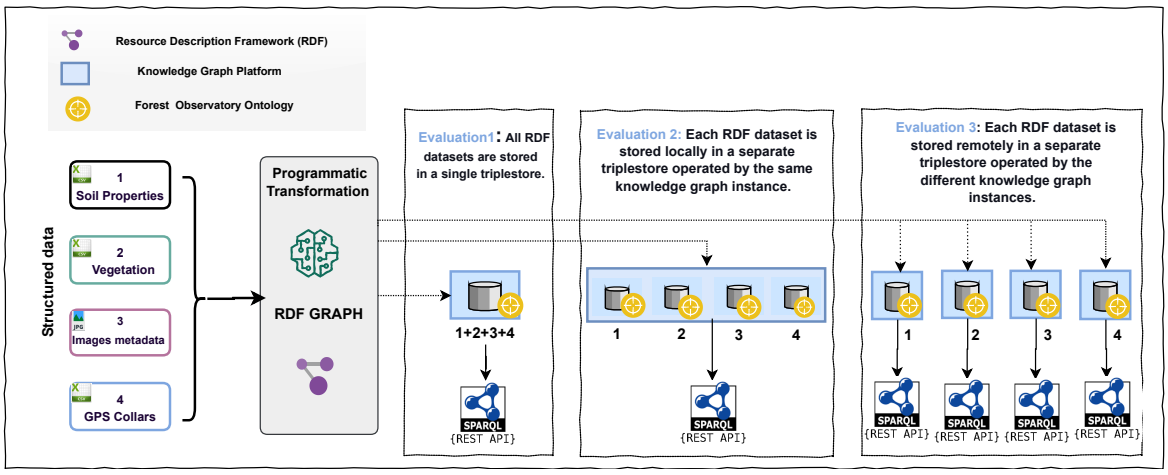


Figure 5.8 Experimental Setup

sensor data observations in JSON format (depending on the specific experiment). Then I transformed JSON data into Resource Description Framework (RDF) graph and loaded graph data into a triple-store database. This meant our algorithm could run and update concurrently or partially by modules. For example, the user may run the algorithm's first part to generate JSON data and then transform the output JSON file into RDF using the second part as an independent code. Our experience found that running the algorithm in stages consumed less time when generating many data observations (e.g., 1000K+). Code Snippet 1 explains the technique used. The UDEO and datasets were connected to create a knowledge graph. Stardog was employed [2], a knowledge graph platform that integrates heterogeneous and isolated data sources. Stardog hosts the triple-store databases and has an IDE (Stardog Studio) capable of performing numerous operations, including SPARQL, GraphQL, artificial intelligence, and machine learning. Complex SPARQL queries were written to test the efficiency of retrieving information and inferring new phenomena. Further details are in the evaluation section.

5.4 Evaluation

The characteristics exhibited in Figure 2 (d), such as sunny days and weekdays, have been modeled using SWRL rules. To illustrate the assessment, the rule were inserted into the database. A query was executed that conveyed, for instance, the condition of a sunny day, assuming a decrease in bike availability. To evaluate the knowledge graph's performance in terms of utility and response time, I evaluated it using three different architectures, each of which differs in how it stores and queries data. In the first instance, Code Snippet 2 was

Code Snippet 2: Data Generate, Transform and Load

```
Part 1– Generate
Function GenerateModel():
    Function GenerateId(size, chars + string):
        return join (chars for i in range(size));
        model = DataModel(Id ,type);
        n = name ;
        v = value ;
        model.add(n,v) ;
    return model ;
x = observation number (integer only) ;
for i in range(x) do
    data= GenerateModel();
    JSON.dump(data);
end for;
Part 2 –Transform ;
Data = ReadJSON(data) ;
for i, r in Data: do
    RDFData=Graph.add(subject,property,object);
    Graph.serialise(RDFData,format=turtle);
end for;
Part 3 – Load ;
StardogConnection=(endpoint,username,password);
DatabaseName = NewDatabase(RDFData) ;
connection = ConnectStardog(RDFData, StardogConnection);
connection.add(RDFData, format=turtle);
connection.commit
```

used to generate three incremental series of synthetic datasets. The number of generated observations was equal for each one. However, the volume varied when serialised into RDF graphs. Table 5.3 shows the kilobyte (KB) size for the generated RDF graphs. Furthermore, Figure 5.9 compares the volume of each data model. It was evident that the Noise Level dataset with more than 1K triples is the largest. That is due to the increased amount of metrics (e.g., CO, NOx) in a single observation compared to other data sources. The objective is to estimate how much data can be stored on a disk for each data model. It helps to understand how much data storage is required for a given use case, depending on the frequency of the observations captured and the number of metrics modeled within each observation. The data stores were interrogated to answer Competency Questions (CQs) such as *where can I park and ride?*. We developed complex SPARQL queries and executed them across the databases in question. Then, Semantic Web Rule Language (SWRL) was inserted into the databases and reran the queries. Each approach was assessed by checking the correctness of each query's result (i.e., the answer to the competency question) and comparing response time on databases with and without SWRL. SPARQL query response time is critical because it directly impacts the usability and performance of applications relying on semantic data. Response time is critical as it directly impacts the completeness and consistency of query responses. Incomplete or incorrect queries may fail to produce results, and query outputs must be manually checked and verified to ensure accuracy. Moreover, when queries take too long—especially in large ontologies or complex scenarios—it can degrade user experience, hinder real-time decision-making, and reduce the system's overall efficiency. Subsequently, I reflected on the impact of complex queries and reasoning processes on data, highlighting some strengths and weaknesses. Furthermore, the collected data were analysed to compare the results and determine if response time is faster after inserting SWRL rules. In other words, to examine if reasoning and setup reduce the query response time. Finally, key findings suggested the most suitable approach for the data marketplace. The accumulated response time observations were analysed with the following steps to detect the difference between the two groups: (i) line graph to visualise and compare the data. (ii) testing the data distribution with the Shapiro-Wilk normality test to determine the appropriate statistical test for detecting the difference between two groups (i.e., parametric or non-parametric). Here, the histograms and test p-value suggested that the observations were not found to be not normally distributed. Accordingly, non-parametric tests were conducted. (iii) using Kruskal-Wallis to check if the two groups of observations (No-rule-SWRL) are related (i.e., sampled from the same distribution). (iV) finally, Mann-Whitney, the statistical test proved if there is a significant difference between the two groups. In other words, to examine if SWRL increased or reduced the query efficiency and response time.

Therefore, the claimed hypotheses can be formulated as:

Hypothesis 1 (H0)**.** *SWRL reduces query response*

Hypothesis 2 (H1)**.** *SWRL does not reduce query response time.*

The significance level of

$$\alpha = 0.05$$

We retain the null hypothesis if the p-value is greater than the significance level.

Data Model	1K	10K	100K
Air Quality	90	903	9023
Beaches	108	859	8580
Docking Stations	119	1188	11877
Museums	153	1215	12141
Noise Level	135	1347	13471
Parking Status	99	918	9180

Table 5.3 Data Model RDF Graph Size (KB)

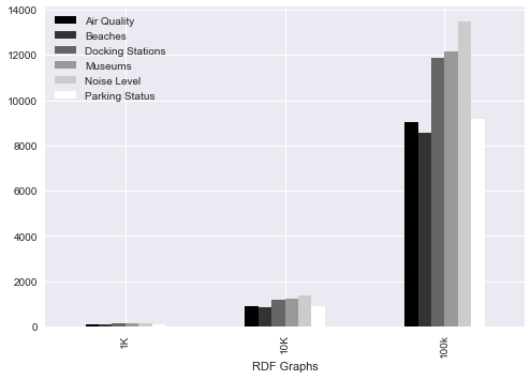


Figure 5.9 Comparison between Data Model Sizes(KB)

5.4.1 Rule-based Reasoning

In the context of the Semantic Web in general and RDF graphs in particular, reasoning, also referred to as inferencing, derives a new phenomenon from a given dataset based on named axioms, applicable rules, and definitions in the data model. Reasoning rules are declarative and represent proven knowledge or concepts modelled by experts. Rule-based inferring uses conditional IF-Then entailment rules. The logical consequences in the IF clause are inferred in the statement of Then. For example, outdoor activities are busier than usual on sunny days. Reasoning can reshape and align data, creating new views of data and connections. More importantly, it validates domain modelling and detects violations. One of the features is the reasoning at query time. Besides the excellent performance, it allows users to specify the type and pay only for its reasoning usage. Reasoning can be enabled or disabled via a simple boolean command. When enabled, rules or axioms are triggered, and reasoning executes according to its value in the database. In this case study, a rule that assumes low rental bike availability during sunny days was created. This assumption was sketched to prove a concept that may not reflect objective reality. As discussed in the coming sections,

the sunny days rule is injected into our database and used to evaluate its performance in three different scenarios.

5.4.2 Evaluation One

The first experiment loaded all six datasets and our ontology (UDEO) into a single database. SPARQL query scripts were executed at Code Snippet 5.3 and 5.4 a hundred times consecutively and respectively. The queries yield vacant car parking spots and available bikes near a geographical location. The time taken by each script was recorded in milliseconds (ms). The experiment was repeated after inserting SWRL, which could infer a sunny day. The SPARQL queries were executed 100 times before and after reasoning (i.e., inserting SWRL). Figures 5.10 and 5.11 show the query responses for each query type (i.e., No-rule and SWRl), and the histograms in Figures 5.12 and 5.13 unveil the data distribution. The obtained p-values from Shapiro-Wilk, Kruskal-Wallis and Mann Whitney, as listed in Table 5.25, were far below the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that SWRL does not reduce response time in this query setup.

Listing 5.1: SunnyDays Rule

```
Prefix rule: <tag:stardog:api:rule:>
[] a rule:SPARQLRule ;
rule:content """
PREFIX␣fiware:
<https://uri.fiware.org/ns/data-models#>
IF
␣{?id␣a␣fiware:BikeHireDockingStation;
␣␣␣␣␣␣␣␣fiware:AvailableBikeNumber;
␣␣␣␣␣␣␣␣?AvialableBikeNumber;
BIND(xsd:integer(AvialableBikeNumber)␣<5
AS␣?SunnyDays)}
THEN
␣{?id␣␣fiware:SunnyDays␣␣?SunnyDays}""".
```

Listing 5.2: PREFIXES for *"Where can I park my car and ride a bicycle?"*

```
PREFIX fiware://<https://uri.fiware.org/ns/data-models#>
PREFIX ngsi: <https://uri.etsi.org/ngsi-ld/>

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

PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX geof:<http://www.opengis.net/def/function/geosparql>

PREFIX unit: <http://qudt.org/vocab/unit#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

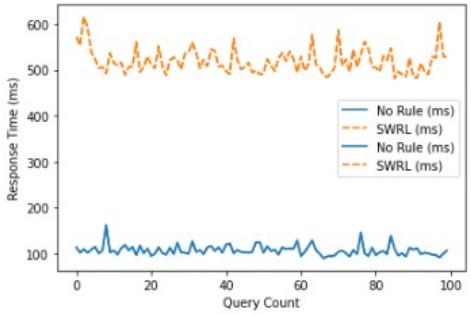


Figure 5.10 Evaluation 1 visualisation

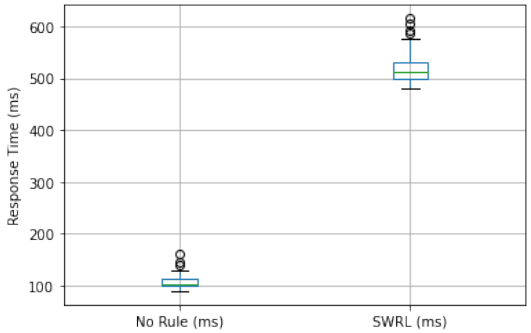


Figure 5.11 Evaluation 1 boxplot

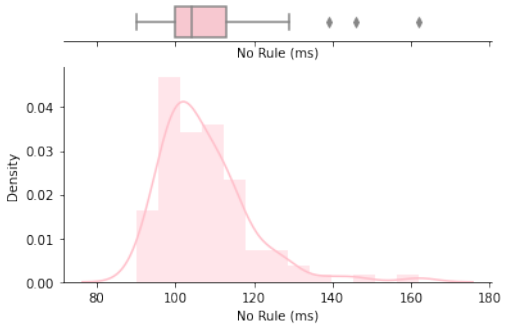


Figure 5.12 Evaluation 1 No-Rule distribution plot

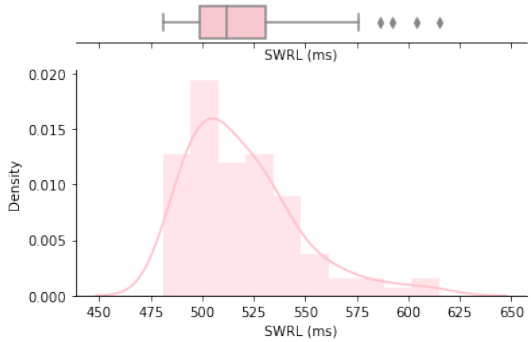


Figure 5.13 Evaluation 1 SWRL distribution plot

Listing 5.3: Evaluation One NoRule Result Response Time = 237 ms

```
Select *
{?id a
fiware:ParkingSpot;
        fiware:category ?category;
        fiware:dataProvider ?dataProvider;
        ngsi:status ?status;
        ngsi:location ?location;
        ngsi:parkingPoint ?ParkingPoint.
?id2 a
fiware:bikeHireDockingStation;
        fiware:availableBikeNumber ?availableBikeNumber;

schema:address ?address;
        ngsi:status ?Bikestatus.
sosa:PointID a pos:Point;
            pos:SOSAPoint
?SOSAPoint.
BIND (geof:distance
    (?ParkingPoint, ?SOSAPoint, unit:Kilometer)
    as ?Distance).
FILTER(xsd:integer(?Distance < 500))
FILTER(REGEX(?Bikestatus, "free"))
FILTER(REGEX(?availableBikeNumber, "1"))
FILTER(REGEX(?category, "offstreet")).}}
LIMIT 1
```

Listing 5.4: Evaluation One-SWRL-Result Response Time= 571 ms

```
Select *
{?id a  fiware:ParkingSpot;
        fiware:category ?category;
        fiware:dataProvider ?dataProvider;
        ngsi:status ?status;
        ngsi:location ?location;
        ngsi:parkingPoint ?ParkingPoint.
?id2 a  fiware:bikeHireDockingStation;
        fiware:availableBikeNumber ?availableBikeNumber;
        schema:address ?address;
        ngsi:status ?Bikestatus;
        fiware:sunnyDays ?SunnyDays.
sosa:PointID a pos:Point;
            pos:sOSAPoint
?SOSAPoint.
BIND (geof:distance
    (?ParkingPoint, ?SOSAPoint, unit:Kilometer)
    as ?Distance).
FILTER(xsd:integer(?Distance < 500))
FILTER(REGEX(?Bikestatus, "free"))
FILTER(REGEX(?availableBikeNumber, "1"))
FILTER(REGEX(?category, "offstreet")).}}
LIMIT 1
```

Listing 5.5: Evaluation Two NoRule Result Response Time= 32 ms

```
SELECT *
{SERVICE <db://BikeHireDockingStation100k>
    {?id a
fiware:BikeHireDockingStation;
            fiware:availableBikeNumber ?availableBikeNumber;
            schema:address ?address;
            ngsi:status ?Bikestatus.}
{SERVICE <db://Parking>
    {?id2 a fiware:ParkingSpot;
            fiware:category ?category;
            fiware:dataProvider ?dataProvider;
            ngsi:status ?status;
            ngsi:location ?location;
            ngsi:ParkingPoint ?ParkingPoint.
 sosa:PointID a pos:Point;
 pos:SOSAPoint  ?SOSAPoint.
 BIND (geof:distance
     (?SOSAPoint,?ParkingPoint, unit:Kilometer)
    as ?Distance).
 FILTER(xsd:integer(?Distance < 290))
 FILTER(REGEX(?Bikestatus, "free"))
 FILTER(REGEX(?availableBikeNumber, "1"))
 FILTER(REGEX(?category, "offstreet")).}}}
 LIMIT 1
```

Listing 5.6: Evaluation Two-SWRL Result Response Time= 47 ms

```
SELECT *
{SERVICE <db://Parking>
{?id2 a
fiware:ParkingSpot;
        fiware:category ?category;
        fiware:dataProvider ?dataProvider;
        ngsi:status ?status;
        ngsi:location ?location.}
        # ngsi:ParkingPoint ?ParkingPoint.
    {?id a fiware:BikeHireDockingStation;
        fiware:availableBikeNumber ?availableBikeNumber;
        fiware:AvailableBikeNumber ?AvialableBikeNumbe;
        schema:address ?address;
        ngsi:status ?Bikestatus;
        fiware:SunnyDays
?SunnyDays.
    sosa:PointID a pos:Point;
    pos:SOSAPoint  ?SOSAPoint.
    # BIND (geof:distance
    #?SOSAPoint, ?ParkingPoint, unit:Kilometer)
    #as ?Distance).}
    # FILTER(xsd:integer(?Distance < 290))
    FILTER(REGEX(?Bikestatus, "free"))
    FILTER(REGEX(?availableBikeNumber, "1")}}
    LIMIT 1
```

5.4.3 Evaluation Two

The second experiment stored each dataset with the UDEO locally but in separate databases using the same Stardog instance. Here, the SPARQL federation answered the competency question. For instance, the query targeted the data source internally by using the SERVICE keyword with the URI typed as *db://database* instead of the SPARQL endpoint URL. For further analysis, the SPARQL queries were executed shown in the Code Snippet 5.5 and 5.6 hundred times, with and without SWRL-recording observations.

In the same manner, the line plot in Figure 5.14 and the box plot in Figure 5.15 compared the two variables. whilst the histograms in Figures 5.16 and 5.17 suggest that the observations do not follow the normal distribution. Noticeably, this setup (locally separated databases) has a quicker response time than the unified database in evaluation one. Similar to evaluation one 5.4.2, the p-values from Shapiro-Wilk, Kruskal-Wallis and Mann-Whitney, as listed in Table 5.25, were less than the significance level of 0.05. Once again, we reject the null hypothesis and recommend that injecting SWRL into the internally distributed databases does not accelerate response time.

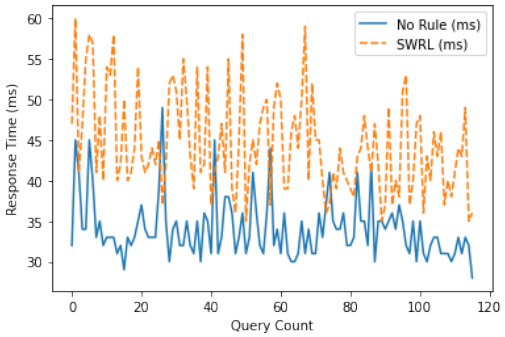


Figure 5.14 Evaluation 2 visualisation

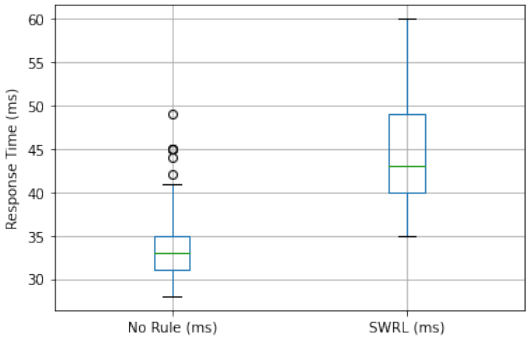


Figure 5.15 Evaluation 2 boxplot

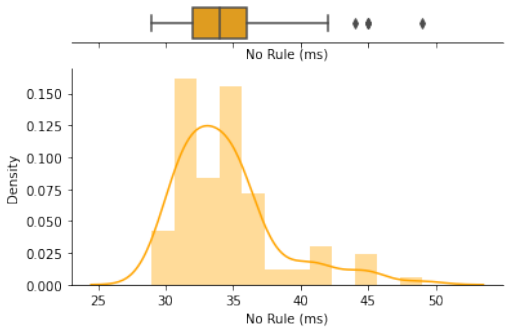


Figure 5.16 Evaluation 2 No-Rule distribution plot

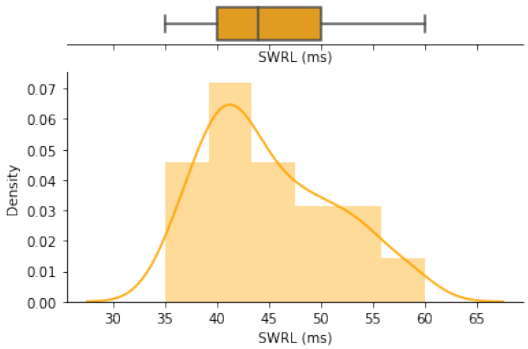


Figure 5.17 Evaluation 2 SWRL distribution plot

5.4.4 Evaluation Three

In the last experiment, each database was stored on a separate computer node under different Stardog instances. Every machine acted as an independent data provider. Federated SPARQL queries were executed, as illustrated in Code Snippet 5.7 and 5.8 to answer our competency question from the desired database without moving or copying data. This time, the query targeted a SPARQL endpoint on a remote machine. Therefore, an IP address was required along with the port number to reference the SERVICE Keyword. Nevertheless, HTTP authentication was also necessary to access the reference SPARQL endpoint. It can be achieved by disabling Stardog security on startup or storing password credentials in the Stardog directory. The same query for all other evaluations was executed concurrently, in the same manner, with and without SWRL. The line plot in Figure 5.19 and the box plot in Figure 5.20 compared the two variables. whilst the histograms in Figures 5.21 and 5.22 suggest that the observations do not follow the normal distribution. Astonishingly, the average response time with SWRL was faster than no-rule. Accordingly, we retain the null hypothesis and conclude that SWRL reduces the query response time in the decentralised data storage manner.

Listing 5.7: Evaluation Three NoRule Result Response Time = 99 ms

```
SELECT *
 {SERVICE <http://192.168.0.128:5820/Parking/query>

     {?id2 a fiware:ParkingSpot;
           fiware:category ?category;
           fiware:dataProvider ?dataProvider;
           ngsi:status ?status;
           ngsi:location ?location.}
{SERVICE <http://192.168.0.128:5820/Bike/query>

    {?id a  fiware:BikeHireDockingStation;
           fiware:availableBikeNumber ?availableBikeNumber;
           fiware:AvailableBikeNumber ?AvialableBikeNumber;
           schema:address ?address;
           ngsi:status ?Bikestatus;}}
LIMIT 1
```

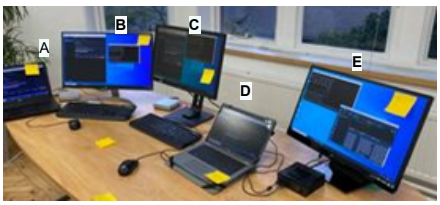
Listing 5.8: Evaluation Three SWRL Result Response Time = 86 ms

```
SELECT *
{SERVICE <http://192.168.0.128:5820/Parking/query>
{?id2 a fiware:ParkingSpot;
       fiware:category ?category;
       fiware:dataProvider ?dataProvider;
       ngsi:status ?status;
       ngsi:location ?location.}

{?id a  fiware:BikeHireDockingStation;
       fiware:availableBikeNumber ?availableBikeNumber;
       fiware:AvailableBikeNumber ?AvialableBikeNumber;
       schema:address ?address;
       ngsi:status ?Bikestatus;
       fiware:SunnyDays  ?SunnyDays.}}
LIMIT 1
```

5.4.5 Results

Data aggregated (n=100) for each evaluation was fed to a Python code for visualisation, exploration, and statistical testing. Figure 5.23 are line graphs that visualise the flow of each trail. Querying the database after inserting SWRL took more time than querying the database without a rule. A descriptive analysis was performed to understand the data. Then, Shapiro-Wilk, Kruskal-Wallis, and Mann-Whitney statistical tests were used to check the distribution of the samples and detect and compare differences between the two independent samples



Name	Data Provider	CPU	SSD (GB)	RAM (GB)
A	Laptop 1	Intel® Core™ i7-6Gen@2.60GHz(8CPUs)	480	16
B	MiniPc3	Quad-Core Intel Atom Z8350 @1.44 GHz	128	8
C	MiniPc2	Quad-Core Intel Atom Z8350 @1.44 GHz	128	8
D	Laptop2	Intel Core i7 1.8GHz 10Gen	512	16
E	MiniPc1	Intel® Core™ i3 5Gen@2.00GHz (4CPUs)	60	8

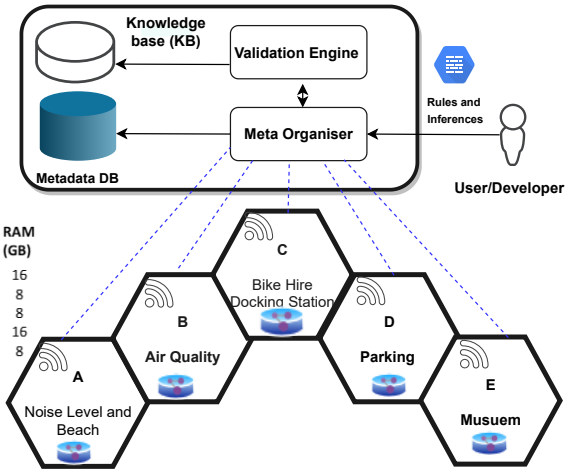


Figure 5.18 Evaluation Three Setup, A to E represent the different data providers' machines. Data providers may store their data on edge using mini pcs or dedicated computers. Here, the user/developer places a request. It gets verified by the validation engine and passed to the meta organiser that knows the data provider that has the requested information.

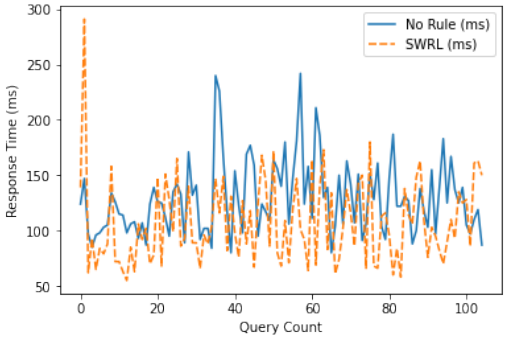


Figure 5.19 Evaluation 3 visualisation

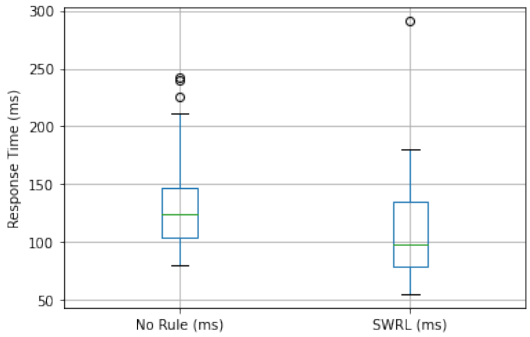


Figure 5.20 Evaluation 3 boxplot

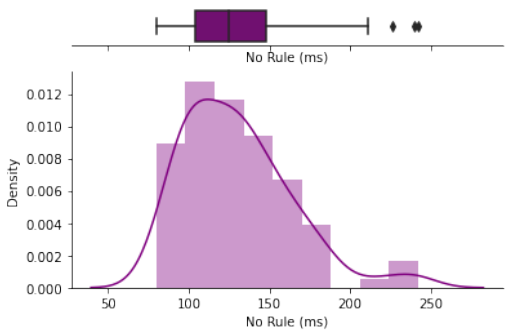


Figure 5.21 Evaluation 3 No-Rule distribution plot

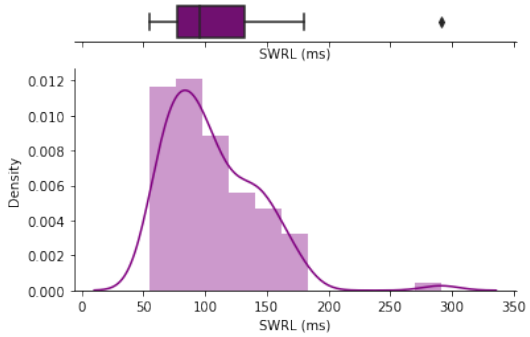
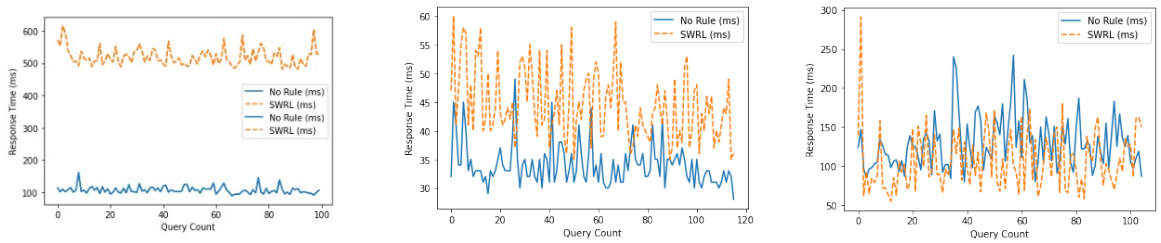


Figure 5.22 Evaluation 3 SWRL distribution plot

(no-rule/SWRL) for each experiment, respectively. *Descriptive Analysis* in Figures 5.24 explored the datasets to help understand the data content and characteristics. For example,



(a) (b) (c)

Figure 5.23 Compare the three evaluations of query response time (ms), with and without reasoning rules inserted. (a) Evaluation One (all datasets and the ontology are stored in a single database). (b) Evaluation Two (each dataset and ontology are stored locally in separate databases). (c) Evaluation Three (each dataset and ontology stored remotely with autonomous data providers)

the line graphs for evaluation one indicated a considerable difference between query time responses for the dataset with and without SWRL. Unlike evaluation two, where the gap narrowed dramatically, it was surprising that querying with SWRL was slightly faster in evaluation 3. This result could be due to the distribution of the datasets over disparate nodes and the remote access, which distinguished evaluation three from the others. The mini-PC nodes that hosted the datasets had relatively small disk spaces ranging from 32GB to 128 GB and no less than 8GB of RAM. Noticeably, the number of observations was equal in all three experiments, the mean fluctuated between approximately 34.45 and 519.13, and the standard deviation was at its lowest point of 3.78 in evaluation two. To further discover the data distribution, the histograms and *Shapiro-Wilk* decided the most appropriate statistical tests. It was clear from the histograms that the data in the three evaluation datasets did not resemble bell curves. Shapiro-Wilk test confirmed the non-normality further through the p-values (<0.05). Therefore, we rejected H0 with a 95 percent confidence interval and concluded that all datasets do not follow the normal distribution. *Kruskal-Wallis test* was the non-parametric test that suited our data distribution. It suggested that the two samples (no-rule/SWRL) for each experiment came from different distributions with p-values of less than 0.05. Figure 5.25 explained in detail the hypotheses result of the evaluations based on their p-values, and the final test, *Mann-Whitney*, was used to determine if the response time was different after inserting SWRL. The result also suggested a significant difference between the samples of each experiment (no-rule/SWRL). Furthermore, the average query response time (in milliseconds) for the three evaluations is presented in Figure 5.26.

	No Rule (ms)	SWRL (ms)
count	100.00	100.00
mean	107.40	519.13
std	11.44	27.17
min	90.00	481.00
25%	100.00	498.75
50%	104.00	512.00
75%	113.00	530.75
max	162.00	615.00

(a)

	No Rule (ms)	SWRL (ms)
count	100.00	100.00
mean	34.45	45.00
std	3.78	6.28
min	29.00	35.00
25%	32.00	40.00
50%	34.00	44.00
75%	36.00	50.00
max	49.00	60.00

(b)

	No Rule (ms)	SWRL (ms)
count	100.00	100.00
mean	130.04	105.68
std	33.95	37.55
min	80.00	55.00
25%	104.00	77.50
50%	125.00	96.00
75%	147.75	131.75
max	242.00	291.00

(c)

Figure 5.24 Descriptive analysis for the three evaluations of query response time (ms)—with and without reasoning rules inserted. Tables explore the datasets' nature and summarise their contents.

Evaluations	Shapiro-Wilk H0 : Data follow a normal distribution. H1 : Data do not follow a normal distribution.				Kruskal-Wallis H0 :Two samples are related H1 :Two samples are not related		Mann Whitney H0 : Sample distributions are equal H1 :Sample distributions are not equal	
	P-value = 0.05							
	No Rule	Result	SWRL	Result	No Rule /SWRL	Result	No Rule /SWRL	Result
One	4.70508E-08	*Reject H0*	5.02E-06	*Reject H0*	0.000	*Reject H0*	0.000	*Reject H0*
Two	3.52513E-08	*Reject H0*	0.0006118	*Reject H0*	0.000	*Reject H0*	0.000	*Reject H0*
Three	4.14385E-05	*Reject H0*	4.94E-07	*Reject H0*	0.000	*Reject H0*	0.000	*Reject H0*

Figure 5.25 Evaluation Statistical Analysis

Evaluations	Average Query Time (ms)				
	SPARQL Query		Difference	%	
	No Rule	SWRL			
One	107.4	519.13	411.73	383	↑
Two	34.45	45	10.55	31	↑
Three	130.04	105.68	-24.36	-19	↓

Figure 5.26 Evaluations query average time comparison

5.4.6 Cost of Adapting Linked Data

The feasibility of linking distant, semantically modeled data sources and reasoning over them using SWRL has been established. It facilitates quick access to diverse data sources. Therefore, data owners can retain and manage their data whilst sharing only their metadata with the IoT data marketplace. The results indicate that semantic data sources efficiently send small packets through the communication network. For instance, when reasoning is performed, the query response time decreases. However, the information overhead and implementation cost must be evaluated before system deployment.

- Typically, information overhead stems from the data structure, runtime, and data exchange. In our methodology, the semantic data sources are database engines whose data are represented in RDF and queried via their SPARQL endpoints. The most significant expense is the creation of an ontology for each data source, as query resolution necessitates a complete merging ontology. Here, ontologies are created by a large consortium of academics and industry professionals [132**?**]. Occasionally, different consortiums may have differing views on the data type, making it difficult to model the data in a single format. Simultaneously, Semantic Web technologies enable the merging and reasoning over ontologies, allowing different classes, for instance, to be defined as equivalent.

- Costs associated with the system implementation may be related to the need for dedicated computers for data storage and retrieval. As depicted in Figure 5.18, These computers had between 8GB and 16GB of RAM, and the smallest solid-state disc (SSD) capable of storing the relevant data source was 60GB in size. Table 5.3 illustrates how much RDF-formatted data can be stored on a single storage disc. For example, an air quality sensor that generates one observation per minute would require approximately one KB of storage space on a disc for 90 observations. Thus, the data provider can determine the amount of information stored on a 60GB SSD. As mentioned, calculating the overhead could determine the total cost before implementation. Although using inexpensive computing devices contributes to the efficiency of this system, the cost may be increased by the labour time of human resources.

5.4.7 Use Cases

Our proposed on-demand data model is applicable in various industries. It offers practical and cost-effective solutions for SMEs. Businesses can build various innovative services enriched with machine learning and AI that respond to end-users personally. In comparison, legacy data trading constrained businesses to acquire bulky datasets that may incur more charges and require high maintenance. (i.e., filtering to process relevant data records). This study's hypothetical use cases concern the tourism and housing industries. The former is a small company that enables consumers to browse and book trips to local attractions, promoting sustainable travel. The latter is a state agency recommending properties with considerably clean air features (i.e., properties in less polluted and quiet areas).

Use Case 1

TripRecomender is a small business that enables end-users to plan green trips to local attractions. It predominantly aims for high-rated customer satisfaction by leveraging AI's self-learning competencies. The company adopted the online chat application, called bots or chatbots. An intelligent program relies on actual data to carry out a specific task. Here, TripRecommender is planning to train its chatbot to search for and suggest local tourist destinations and sustainable travel solutions, increasing customer satisfaction and boosting the company's revenue. Well-trained bots can reduce the time and effort spent on monotonous trip planning. TripRecommender considers users' requests and suggests tailored options based on previously known information. The company's chatbot requires sufficient relevant data to analyse and turn into meaningful information to accomplish this task. Relevant data records may exist in separate data sources, making aggregating challenging. With our on-demand data model, TripRecommender can query and retrieve granular data records that are fit to train the bot. (e.g., ten years of data for either the local beaches occupation rate or the availability of rental bikes on weekends at the local docking stations). Subsequently, it recommends customised offers and infers new events. For example, the SPARQL query in **Code Snippet 5.9** expresses how our model retrieves certain weekend information about (i) a local beach's name, services, and occupancy rate, (ii) a museum's opening hours, and (iii) an available rental bike location.

Listing 5.9: Use Case 1 - TripRecommender

```
PREFIX : <http://api.stardog.com/>
PREFIX stardog: <tag:stardog:api:>
PREFIX schema: <https://schema.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ngsi-ld:<https://uri.etsi.org/ngsi-ld/default-context/>
PREFIX fiware:<https://uri.fiware.org/ns/data-models#>
PREFIX ngsi: <https://uri.etsi.org/ngsi-ld/>
SELECT *
    {?id a fiware:Beach;
    ngsi:name ?Name;
    fiware:facilities ?Facilities;
    fiware:occupationRate ?OccupationRate;
    fiware:Weekends    ?Weekends.
{SERVICE   <http://192.168.0.78:5820/Museum/query>
    {?id2 a fiware:Museum;
    fiware:openingHoursSpecification   ?OpeningHours.
{SERVICE   <http://192.168.0.128:5820/Bike/query>
    {?id3 a fiware:BikeHireDockingStation;
    fiware:availableBikeNumber ?availableBikeNumber;
    schema:address  ?address;
    ngsi:status ?Bikestatus. }}}}}
```

Use Case 2

CleanAir Housing is a medium-sized business that sells and rents out residential properties. Recently, the company noticed a staleness and price drop for properties in highly polluted areas based on its sales records. Conversely, homes in quiet and less polluted areas will most likely sell in a year. Air pollution could negatively impact health. It happens when particular gases and liquid particles are released into the atmosphere, forming PM2.5 and PM10 particles and elevating levels of carbon monoxide (CO) and nitrogen dioxide (NO2) pollutants above the clean air legal limit. Sources of these toxic gases include vehicle exhaust, factories, and domestic combustion. Measuring the Air Quality Index (AQI) can assess the air quality in areas of interest. As a result, CleanAir Housing decided to integrate AI-powered services to (i) predict sales forecasts based on historical data and air pollution levels, (ii) recommend the optimal price to match the expected value, and (iii) hunt for local fast-selling homes. Several environmental data records are needed to calculate air quality and noise levels. Traditional ways to acquire such data are by deploying thousands of sensors or purchasing bulky ecological datasets. Both options demand time and funds. Hence, further data mining, processing, and analysis are required to extract valuable insights. Integrating our on-demand data model enables the filtering and extraction of the needed metrics from multiple datasets, forming search data that are fit to train the company's AI. For instance, the SPARQL query script in **Code Snippet 5.10** combined metrics from air quality and noise level datasets to filter out the addresses in areas with low noise levels and good AQI.

Listing 5.10: Use Case 2 - CleanAir Housing

```
SELECT*
{SERVICE  <http://192.168.0.128:5820/AirQuality/query>;
    {?id a  fiware:AirQualityObserved;
        schema:address  ?address;
        fiware:dateObserved ?date;
        fiware:Precipitation ?Precipitation;
        fiware:Reliability  ?Reliability;
        fiware:WindDirection?WindDirection;
        fiware:AirQualityIndex ?AirQualityIndex;
        ngsi-ld:Co  ?Co;
        ngsi-ld:CO_Level ?CO_Level;
        ngsi-ld:No?No;
        ngsi-ld:Nox ?Nox;
        ngsi-ld:No2 ?No2;
        ngsi-ld:So2 ?So2.
 {SERVICE  <http://192.168.0.78:5820/Noise/query>
    ?id2 a  fiware:NoiseLevelObserved;
        fiware:DateObservedFrom ?DateObservedFrom;
        fiware:DateObservedTo ?DateObservedTo;
        fiware:frequencies ?frequencies;
        fiware:DataProvider ?DataProvider;
        ngsi:location  ?location;
        ngsi-ld:lAeq_d ?lAeq_d;
        ngsi-ld:lAmax ?lAmax.
        FILTER(xsd:float(?lAmax) > 0.72)
        FILTER(REGEX(?address, "A"))
        FILTER(REGEX(?CO_Level,"Low"))
        FILTER(xsd:integer(?AirQualityIndex) < 100)
        FILTER(xsd:integer(?Nox) < 100)}}}
    Limit10
```

5.5 Discussion

SynchroniCity data marketplace sells sensor data in bulk. Consumers interested in specific observations from different sensors (e.g., air quality and noise level) must purchase each sensor's dataset separately. Such a practice may incur more charges and cause a high-latency network. Our study extended SynchroniCity data marketplace with Semantic Web technologies to allow consumers to pay for the sensor's information they need - instead of buying the entire dataset. Consumers can acquire multiple observations from various data providers to fulfill their orders. For example, finding nearby parking spaces and museum opening hours on a particular date and time.

Our end-product consists of a user-friendly interface with an interactive map and a semantic data model. More specifically, (i) a novel semantic model was proposed. It encompassed an Urban Data Exchange Ontology (UDEO) and FIWARE synthetic datasets for six different providers. (ii) three different experiments were conducted, as shown in 5.18 to determine

the most practical modelling and storage solutions for the IoT data marketplaces. (iii) the experiments were evaluated to demonstrate the effectiveness of the semantic modelling and SWRL - using different SPARQL queries to answer related competency questions. The evaluation results support the hypothesis that reasoning over distributed data sources could be the ideal architecture for the IoT data marketplace. Evidently, in evaluation one 5.4.2, querying after inserting rules took a long time, and the time-lapse between queries with and without rules is relatively slow. In evaluation two 5.4.3, the query time response was dramatically reduced compared to evaluation one. It is worth mentioning that to make SWRL rules work, the rule was inserted independently in each database, unlike the evaluation one, where dealing with a single database. Even when SWRL is inserted in each database, it did not activate with the SERVICE keyword. The query line had to be executed within the rules inserted in the database. Query line targeting rules and reasoning responded when calling the database internally in the Stardog Studio workspace. In evaluation three 5.4.4, the requested data can be obtained remotely via HTTP, using the host's IP address, port, and database name. The user query breaks into triple patterns that interrogate data sources SPARQL endpoints for results. Figure 5.26 compared the average evaluation time between querying databases with and without SWRL. Evaluation One showed the highest difference in query response time among the evaluations. Unexpectedly, the average response time on SWRL databases was lower than without rules. Therefore, valuable insights can be drawn from our semantic model results as follows:

Semantic modelling and reasoning: Extracting explicit information from IoT SynchroniCity datasets was challenging since these data lacked formal definitions for widely shared standards. Our semantic model transformed them into queriable triplestores. The adapter (code) mapped the data to RDF whilst referencing the UDEO. Abstract rules were inserted into the databases to trigger reasoning such as *Sunnydays and weekends*. Sunny days rule sets available rental bikes level low, assuming higher demand on such days. whilst the weekends' rule deduced high occupancy rates on local attractions such as beaches and museums. Reasoning quickened query response time in experiment three by reducing the search space whilst filtering out information adhering to the rules. SPARQL queries retrieved granular and semantically enriched and reasoned information from different datasets, stored locally and remotely. As a result, customised data requests can be achieved at low costs.
Interoperability: It's suggested that experiment three's approach is interoperable. SPARQL allowed remote access through its endpoints, achieving seamless data sharing between different RDF databases stored on heterogeneous machines.
***Edge computing*:** In experiment three, the RDF datasets were distributed on separate comput-

ers operating independently. Executing data on these edge computers satisfied the horizontal scaling property, provided storage capacity, allowed computational flexibility (i.e., semantic modelling), and maintained low network latency (i.e., transmitting query results instead of the whole dataset).

Limitations: Despite that, our semantic model slashed data price, reduced network latency, and cut down information overload in the SynchroniCity data marketplace- yet this approach has some drawbacks. In particular, *the pricing structure, data platform security, data quality, and safe dissemination.* The pricing structure of our model allows consumers to pay for desired information instead of an entire dataset. Although it costs less, working out the total price of an order could be a complex task. Managing diverse data providers with varying data tariffs, broker fees, and applicable taxes presents a significant challenge. Each of these has independent calculations and may change over time. Therefore, offering fixed and competitive charges is an open challenge. Therefore, it's highly recommended to add a self-configuring pricing model that standardises and price-marks data records across independent stores. For example, set one reasonable price for each data record- automatically updating to match the data market's supply and demand, then adjust the broker fees to be a fair percentage of the total bill. Regarding security, accessing and querying data stored in remote machines via HTTP pose risks. Stardog offers security options such as authentication and password encryption; so far, its default security settings are considered minimal for network communications. Therefore, it's recommended using the Secure Sockets Layer (SSL) encryption when deploying Stardog in production mode. Concerning data quality, *Synthetic data* used in this study are consistent with good quality, whilst real sensors data may have errors and missing values. Hence, data quality should be carefully addressed to replicate our real-life study. Accurate machine learning algorithms and artificial intelligence to detect and automatically correct errors are also suggested. The information retrieved to fulfil consumers' requests creates new datasets. These datasets have diverse sources collected by sensors owned by different stakeholders. Thus, publishing them may raise data ownership and privacy concerns. A remedy could be building a tool that (i) traces the data lineage and accurately identifies the owner. (ii) applies a GDPR-compliant privacy policy agreed upon by all parties (data buyer, seller/owner and broker).

5.6 Summary

Data marketplaces are a new category of online marketplaces. Therefore, they are not well-researched within the academic community or well-implemented within the industry. SynchroniCity represents the first attempt to deliver a Single Digital City Market for Europe

by piloting its foundations at scale in 11 reference zones - 8 European cities and 3 more worldwide cities - connecting 34 partners from 11 countries across 4 continents. The primary goal is to meet the data needs of consumers. Data marketplaces also emphasise vital challenges around data acquisition. Data marketplaces incentivise owners to share the gathered data and recover part of the acquisition costs. A fundamental issue of syntactic data marketplaces such as SynchroniCity is that they do not selectively provide a mechanism to buy data. It means data consumers have to buy the entire datasets that data owners offer. FooDS's approach enabled selective querying of annotated IoT data. This approach allowed users to access only the data they needed, reducing costs and improving efficiency.

Most of this chapter has been published in ACM Transactions on Internet of Things as: Naeima Hamed, Andrea Gaglione, Alex Gluhak, Omer Rana, and Charith Perera. 2023. Query Interface for Smart City Internet of Things Data Marketplaces: A Case Study. ACM Trans. Internet Things 4, 3, Article 19 (August 2023), 39 pages.

The contributions are as follows: Conceptualisation—Charith Perera, Andrea Gaglione, Alex Gluhak; Methodology—Naeima Hamed; Software—Naeima Hamed; Validation—Naeima Hamed; Formal analysis—Naeima Hamed; Resources—Charith Perera, Omer Rana, Naeima Hamed; Data curation—Naeima Hamed; Writing (original draft)—Naeima Hamed, Charith Perera; Writing (review and editing)—Naeima Hamed, Charith Perera, Omer Rana; Visualisation—Naeima Hamed, Charith Perera; Project administration—Charith Perera, Andrea Gaglione, Alex Gluhak. All authors have read and approved the published version of the manuscript.

Chapter 6

Conclusion

This doctoral thesis proposed a novel approach for integrating diverse wildlife data and using these integrated data to build deep learning models that predict elephants' geo-locations and poaching likelihood. This research focuses on the situation of Bornean elephants in the Lower Kinabatangan region of Sabah, Malaysian Borneo, which remains a critical concern as they face the persistent threat of poaching, human-elephant conflict and habitat loss. Injuries to humans and elephants sometimes occur due to conflicts near oil palm plantations. Elephants invade the oil palm plantations, causing serious harm to humans, properties and machines. Moreover, elephants often fall victim to snare traps set for wild boar and deer in forest areas like the Kinabatangan floodplain. Since 2010, it's estimated that 20% of Bornean elephants have been injured by these snares. Recent statistics show that the population of Bornean elephants has dropped to less than 1,500, largely owing to poaching driven by illicit ivory trade. In Sabah, boats easily access forests; between 2010 and 2021, at least 200 elephants died in Malaysia, with many incidents linked to poisoning near oil palm plantations. In 2013, 14 Borneo pygmy elephants were found poisoned near the Gunung Rara Forest Reserve, close to logging camps and palm plantations. Similar cases occurred in 2018, when six elephants were poisoned on plantations in Sabah, and in 2019, when three more were killed near palm oil plantations. These incidents highlight the ongoing conflict between elephants and the expansion of agricultural land, particularly for oil palm cultivation. Poaching poses dangers to wildlife officers as well. Rangers who protect wildlife often put their life at risk when confronting armed poachers. This research collaborated with a research and training facility named Danau Girang Field Centre (DGFC), which is managed jointly by the Sabah Wildlife Department and Cardiff University. DGFC was the research hub where most of the interviews with wildlife researchers and discussion groups took place.

6.1 Research Questions and Contributions

The overarching research question is *Can a 'Linked Data Store' be developed to answer questions supporting wildlife research and conservation activities in the wild?*

To reach an answer to the overarching research question, three sub-questions have been asked:

Research Question 1 (RQ1) *Can an effective data management approach be developed to integrate heterogeneous wildlife data from disparate sources?*

RQ1 Contribution Summary: To find an effective data management approach, thirteen Open Data Observatories were selected, examined, and compared. Open Data Observatories are online data platforms that integrate heterogeneous data from disparate sources. The comparison was based on their data types, domain coverage, accessibility, and usability. Their data management approaches were scrutinised to learn from them and find a suitable approach to adopt in this research. The findings from the literature review guided the recommendation to employ semantic web technologies as an effective data management approach for this research. Semantic web technologies have the capability to link and integrate heterogeneous data from disparate sources. Consequently, another review of the literature related to wildlife data management was conducted, focusing on the application of semantic web technologies in modelling wildlife data. Here, the development and advantages of using ontologies and knowledge graphs were briefly explored, respectively. Existing studies on knowledge graphs in predictive modelling and crime prediction were also examined and compared to this approach alongside the advancements in wildlife crime prediction techniques.

Research Question 2 (RQ2) Can a 'Linked Data Store' be developed to answer questions supporting wildlife research and conservation activities?

RQ2 Contribution Summary: To address RQ2, an ontology named the Forest Observatory Ontology (FOO) and its online documentation[1] was developed to standardise wildlife data generated by sensors. The Forest Observatory Ontology (FOO) is a novel ontology built from data collected from wildlife research. It integrates elements from established ontologies to unify the Internet of Things (IoT) and wildlife concepts (biodiversity, conservation biology, habitat fragmentation, and endangered species management). To break wildlife data silos, FOO was populated or instantiated with four heterogeneous datasets transformed into Resource Description Framework (RDF) referencing FOO to produce the Forest Observatory Ontology Data Store (FooDS). To access and use FooDS, an interface was created to enable authorised users to script granular (SPARQL) search queries and retrieve instant answers to questions from integrated and remotely located datasets. RQ2 contribution provides a

[1] w3id.org/def/foo#

novel (modular) approach to link, manage, and analyse wildlife data to answer questions that support conservation and wildlife research.

Research Question 3 (RQ3) *Can prediction models be developed to predict poaching crimes by using the developed 'Linked Data Store'?*

RQ3 Contribution Summary: To build the predictive too named PoachNet, granular data were extracted from FooDS to infer poaching likelihood. PoachNet applied sequential neural network model to predict an Asian elephant's geo-locations. Rule-based semantic reasoning was employed to infer poaching event based on elephant *Seri's* closeness to the oil palm plantation in Sabah forest, Malaysia. PoachNet equips conservationists with a useful tool to predict future elephant locations and poaching likelihood.

Research Question 4 (RQ4): *Can the Linked Data Store's semantic web data management approach be generalised to another domain for various purposes?*

RQ4 Contribution Summary: To answer this research question, FooDS's semantic web data management approach was generalised to the IoT data marketplace. In traditional data marketplaces, data are often sold as entire datasets. This approach can be expensive and inefficient, as consumers may not require the entirety of the dataset but only specific observations or subsets of the data. to address this challenge, an ontology was developed with expert input and reuse, populated with datasets from various sensors to construct knowledge graphs and apply reasoning. FooDS's semantic web data management approach allowed data consumers to acquire granular data records tailored to their needs from various data sources or providers, instead of purchasing the entire datasets. This research experiment was evaluated in three different marketplace setups, measuring key parameters and extracting recommendations for future deployments.

6.2 Research Novelty

This research presents a significant advancement in semantic data integration for wildlife by addressing gaps identified in selected Open Data Observatories and existing wildlife and environmental data integration efforts (e.g., Global Biodiversity Information Facility (GBIF)[2] and World Environment Situation Room (WESR)[3] by UNEP). Whilst selected data platforms provide services by integrating various datasets into a single platform, offering downloadable datasets, visualisations, and analysis options, they do not allow users to query and acquire data at a granular level from different datasets simultaneously. To the best of our knowledge,

[2] gbif.org
[3] wesr.unep.org/

these platforms do not enable users to create on-demand data sets combining records from multiple datasets, which are efficient and cost-effective for training AI models.

This research approach, using the Forest Observatory Ontology (FOO) populated with heterogeneous wildlife datasets creating an ontology-based knowledge graph(s) named the Forest Observatory Ontology Data Store (FooDS), overcomes these limitations. FooDS enables users to query an integrated wildlife dataset published as a URL [4] and obtain data on demand.

FOO was developed through qualitative analysis, including interviews with biologists, ethnographic studies, and discussions with wildlife researchers. FOO was validated using open-source tools, expert evaluations, and practical use cases with Competency Questions (CQs) formulated as SPARQL queries. Whilst other ontologies and platforms provide robust documentation and support, FOO's iterative development process following the LOT methodology ensures it meets the World Wide Web Consortium (W3C) recommendation and the diverse needs of wildlife researchers and conservationists.

FooDS was generalised and applied to a case study in the IoT data marketplace (Chapter 5). An ontology was collaboratively developed with experts, including data scientists, semantic web developers, and practitioners, by reusing the SOSA ontology. This ontology was populated with six heterogeneous sensor datasets to construct knowledge graphs, and semantic web Rule Language (SWRL) reasoning was applied to three use cases to evaluate performance and recommend optimal data configurations for future deployments. The findings demonstrated that SWRL can effectively assert rules within data whilst enabling efficient data storage at the edge (close to the source) and remote access via SPARQL queries.

6.3 Future Work

This research provides a strong foundation for expanding the Forest Observatory Ontology (FOO) to support a broad range of applications. Future work will focus on extending the ontology, integrating AI-driven systems, and enhancing data interoperability to address emerging challenges and opportunities in wildlife conservation.

Expanding the Ontology: FOO can be enriched with additional classes and properties to represent diverse wildlife species and sensor types. For instance, whilst the current ontology includes Asian elephants (Elephas maximus), it could be expanded to include African elephants (Loxodonta africana) and forest elephants (Loxodonta cyclotis), enabling comparative studies across species and regions. Similarly, integrating data types such as crop data (e.g., grass, palm shoots, bananas), water sources, and deforestation metrics would

[4] https://w3id.org/def/fooDS

significantly broaden its scope. Specific Competency Questions (CQs) .3, such as CQ14 ('How did Elephant X's movements change with climate change in 2014?'), CQ58 ('What locations could have snares?'), and CQ99 ('Where are the water sources?'), highlight the need for incorporating weather data, snare locations, and other critical metrics.

Addressing Data Gaps and Environmental Challenges: Field challenges encountered during this research emphasised the need for reliable and protected sensor data collection. Soil condition metrics, such as moisture and fertility, identified in CQs 61 ('What is the soil condition during certain times of the year?') and 76 ('What is the soil moisture in a certain location?'), are essential for answering complex queries. However, interference from wildlife—such as monkeys tampering with soil sensors and elephants damaging air quality sensors—highlighted the importance of developing wildlife-proof sensor solutions.

Leveraging Advanced Technologies: Incorporating datasets from Light Detection and Ranging (LiDAR), air quality, and noise/sound sensors can help address gaps caused by unavailable or inaccessible data during this study. Integrating AI-driven systems such as Large Language Models (LLMs) could facilitate natural language queries, making FOO more user-friendly. Predictive recommender systems could support conservationists in decision-making by identifying optimal interventions based on historical and real-time data. Additionally, connecting predictive models to live data streams from GPS and other sensors could enable real-time alert systems, helping rangers optimise patrol routes and respond proactively to threats.

Privacy, Security, and Collaboration: Data privacy and security are crucial, particularly for sensitive wildlife and geographic information in areas vulnerable to poaching. Future developments should focus on implementing encryption, controlled access, and anonymisation techniques to safeguard these data. Collaboration with universities, conservation organisations, and industry will be key to encouraging widespread adoption of FOO. By fostering community input and establishing clear processes for adding new data, classes, and properties, FOO can remain relevant to the evolving needs of conservation.

Achieving System Interoperability: FOO has been automatically classified by BioPortal as a view of the semantic web for Earth and Environment Technology Ontology (SWEET) (bioportal.bioontology.org/ontologies/SWEET), aligning it with the SWEET ontology framework. Integrating FOO further with platforms like the Global Biodiversity Information Facility (GBIF) and the World Environment Situation Room (WESR) holds the potential to enhance data sharing and aggregation significantly. This interoperability would elevate FOO into a critical tool for global conservation efforts, fostering connections between datasets and empowering researchers and conservationists to collaboratively tackle pressing challenges.

6.4 Concluding Remarks

Wildlife research activities generate vast data on ecosystems and species interactions, often collected from various independent projects. Forest Observatories are online platforms that aggregate, curate, integrate, store, and analyse this data to support effective forest monitoring and answer complex questions. However, integrating data from diverse sources can be challenging due to different data formats and management systems.

A novel solution to this problem involves using knowledge graphs built on ontologies to integrate diverse wildlife data into Forest Observatories. This thesis introduced the Forest Observatory Ontology (FOO), created to link and standardise entities in wildlife research data. FOO was developed through qualitative analysis, including interviews with eight biologists, four ethnographic studies, and discussions with eleven wildlife researchers. FOO reused classes and properties from existing ontologies (W3C recommendation) to standardise FOO's concepts and relationships. The ontology was populated with four semantically modelled wildlife datasets, resulting in the Forest Observatory Ontology Data Store (FooDS)- an ontology-based knowledge graph with over six million data triples. The structure and usability of FOO were validated using open-source tools, expert evaluations, and practical use cases. FOO in turtle format, FOO's documentation and FooDS in turtle format and their resource website are published at *https://w3id.org/def/foo, https://w3id.org/def/fooDocs, https://w3id.org/def/fooDS, and https://ontology.forest-observatory.cardiff.ac.uk*

In conclusion, semantic web technologies, such as ontologies and knowledge graphs, offer significant advantages for data scientists. These technologies enable computers to understand and process data effectively, allowing for automated reasoning, data integration, and complex querying. Integrating data science and advanced analytics in wildlife conservation can revolutionise the field by enhancing predictive capabilities, optimising resource allocation, and accurately measuring conservation efforts. Advanced models can predict poaching incidents, optimise patrol routes, and assess conservation strategies' effectiveness. Despite challenges in wildlife data collection, these technologies provide cost-effective solutions for underfunded conservation programmes, facilitating better resource distribution and strategically deploying rangers and ground truth sensors in high-risk areas.

Bibliography

[1] (2022). Forest Observatory Ontology - Summary | NCBO BioPortal.

[2] (2022). The Enterprise Knowledge Graph Platform | Stardog.

[3] Abram, N., Skara, B., Othman, N., Ancrenaz, M., Mengersen, K., and Goossens, B. (2022). Understanding the spatial distribution and hot spots of collared bornean elephants in a multi-use landscape. *Scientific Reports*, 12(1):12830.

[4] Ahmadi, H., Arji, G., Shahmoradi, L., Safdari, R., Nilashi, M., and Alizadeh, M. (2019). *The application of internet of things in healthcare: a systematic literature review and classification*, volume 18. Springer Berlin Heidelberg.

[5] Akala, V., Ejidike, B., and Olaniyi, O. (2023). Habitat suitability modeling of african forest elephant (loxodonta cyclotis) in omo forest reserve, ogun state, nigeria. *Journal of Research in Forestry, Wildlife and Environment*, 15(2):158–168.

[6] Alfaifi, Y. (2022). Ontology development methodology: A systematic review and case study. In *2022 2nd International Conference on Computing and Information Technology (ICCIT)*, pages 446–450. IEEE.

[7] Alfred, R., Ambu, L., Nathan, S., and Goossens, B. (2011). Current status of asian elephants in borneo. *Gajah*, 35:29–35.

[8] Ali, M., Alexopoulos, C., and Charalabidis, Y. (2022). A comprehensive review of open data platforms, prevalent technologies, and functionalities. In *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, ICEGOV '22, page 203–214, New York, NY, USA. Association for Computing Machinery.

[9] Alkhalil, A. and Ramadan, R. A. (2017). IoT Data Provenance Implementation Challenges. *Procedia Computer Science*, 109(2014):1134–1139.

[10] Almanie, T., Mirza, R., and Lor, E. (2015). Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. *International Journal of Data Mining & Knowledge Management Process*, 5(4):01–19.

[11] Alobaid, A., Garijo, D., Poveda-Villalón, M., Santana-Perez, I., Fernández-Izquierdo, A., and Corcho, O. (2019). Automating ontology engineering support activities with ontoology. *Journal of Web Semantics*, 57:100472.

[12] Aminu, E. F., Oyefolahan, I. O., Abdullahi, M. B., and Salaudeen, M. T. (2020). A review on ontology development methodologies for developing ontological knowledge representation systems for various domains.

[13] Arenas, M. and Pérez, J. (2011). Querying semantic web data with sparql. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 305–316.

[14] Arslan, M., Desconnets, J.-C., and Mougenot, I. (2022). Environmental and life sciences observations in knowledge graphs using nlp techniques to support multidisciplinary studies. *Procedia Computer Science*, 201:543–550. The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40).

[15] Athanasiadis, I. N., Villa, F., Examiliotou, G., Iliopoulos, Y., and Mertzanis, Y. (2014). Towards a semantic framework for wildlife modeling. In *EnviroInfo*, pages 287–292.

[16] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

[17] Bakana, S. R. and Zhang, Y. (2020). Mitigating Wild Animals Poaching Through State-of-the-art Multimedia Data Mining Techniques: A Review.

[18] Bakker, K. (2022). *The sounds of life: How digital technology is bringing us closer to the worlds of animals and plants*. Princeton University Press.

[19] Balmford, A., Gaston, K. J., Blyth, S., James, A., and Kapos, V. (2003). Global variation in terrestrial conservation costs, conservation benefits, and unmet conservation needs. *Proceedings of the National Academy of Sciences*, 100(3):1046–1050.

[20] Bansal, M., Chana, I., and Clarke, S. (2020). A survey on iot big data: Current status, 13 v's challenges, and future directions. *ACM Comput. Surv.*, 53(6).

[21] Barnett, D. T., Adler, P. B., Chemel, B. R., Duffy, P. A., Enquist, B. J., Grace, J. B., Harrison, S., Peet, R. K., Schimel, D. S., Stohlgren, T. J., et al. (2019). The plant diversity sampling design for the national ecological observatory network. *Ecosphere*, 10(2):e02603.

[22] Beek, W., Rietveld, L., Ilievski, F., and Schlobach, S. (2017). Lod lab: scalable linked data processing. *Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering: 12th International Summer School 2016, Aberdeen, UK, September 5-9, 2016, Tutorial Lectures 12*, pages 124–155.

[23] Behnke, J. (2017). Nasa's earth observing system data and information system (eosdis). Technical report.

[24] Bermudez-Edo, M., Elsaleh, T., Barnaghi, P., and Taylor, K. (2016). Iot-lite: A lightweight semantic model for the internet of things. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, pages 90–97.

[25] Berzaghi, F., Bretagnolle, F., Durand-Bessart, C., and Blake, S. (2023). Megaherbivores modify forest structure and increase carbon stocks through multiple pathways. *Proceedings of the National Academy of Sciences*, 120(5):e2201832120.

[26] Bezuidenhout, L. (2020). Being fair about the design of fair data standards. *Digital Government: Research and Practice*, 1(3):1–7.

[27] Bihu, R. (2023). Qualitative data analysis: Novelty in deductive and inductive coding.

[28] Bizer, C., Heath, T., and Berners-Lee, T. (2008). Linked data: Principles and state of the art. In *World wide web conference*, volume 1, page 40.

[29] Blagec, K., Barbosa-Silva, A., Ott, S., and Samwald, M. (2022). A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. *Scientific Data*, 9(1):322.

[30] Blomqvist, E., Hammar, K., and Presutti, V. (2016). Engineering ontologies with patterns-the extreme design methodology. *Ontology Engineering with Ontology Design Patterns*, (25):23–50.

[31] Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, pages 427–434.

[32] Bonatti, P. A., Decker, S., Polleres, A., and Presutti, V. (2019). Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). In *Dagstuhl reports*, volume 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[33] Bonnald, J., Cornette, R., Pichard, M., Asalu, E., and Krief, S. (2023). Phenotypical characterization of african savannah and forest elephants, with special emphasis on hybrids: the case of kibale national park, uganda. *Oryx*, 57(2):188–195.

[34] Bourgeois, D., Bourgeois, A. G., and Ashok, A. (2022). Demo: Ross: A low-cost portable mobile robot for soil health sensing. In *2022 14th International Conference on COMmunication Systems and NETworkS (COMSNETS)*, pages 436–437.

[35] Brum-Bastos, V., Long, J., Church, K., Robson, G., de Paula, R., and Demšar, U. (2020). Multi-source data fusion of optical satellite imagery to characterize habitat selection from wildlife tracking data. *Ecological Informatics*, 60:101149.

[36] Buchelt, A., Adrowitzer, A., Kieseberg, P., Gollob, C., Nothdurft, A., Eresheim, S., Tschiatschek, S., Stampfer, K., and Holzinger, A. (2024). Exploring artificial intelligence for applications of drones in forest ecology and management. *Forest Ecology and Management*, 551:121530.

[37] Bugbee, K., le Roux, J., Sisco, A., Kaulfus, A., Staton, P., Woods, C., Dixon, V., Lynnes, C., and Ramachandran, R. (2021). Improving discovery and use of nasa's earth observation data through metadata quality assessments. *Data Science Journal*, 20:17–17.

[38] Byabazaire, J., O'Hare, G., and Delaney, D. (2020). Data Quality and Trust : A Perception from Shared Data in IoT. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6.

[39] Byrne, D. (2022). A worked example of braun and clarke's approach to reflexive thematic analysis. *Quality & quantity*, 56(3):1391–1412.

[40] Casazza, M. L., Lorenz, A. A., Overton, C. T., Matchett, E. L., Mott, A. L., Mackell, D. A., and McDuie, F. (2023). Aims for wildlife: Developing an automated interactive monitoring system to integrate real-time movement and environmental data for true adaptive management. *Journal of Environmental Management*, 345:118636.

[41] Chatterjee, D. and Rao, S. (2020). Computational Sustainability. *ACM Computing Surveys (CSUR)*, 53(5).

[42] Cheah, C. and Yoganand, K. (2022). Recent estimate of asian elephants in borneo reveals a smaller population. *Wildlife Biology*, 2022(2):e01024.

[43] Chen, W., Zhang, W., Liu, D., Li, W., Shi, X., and Fang, F. (2021). Data-driven multimodal patrol planning for anti-poaching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15270–15277.

[44] Chen, X., Cho, Y., and Jang, S. Y. (2015). Crime prediction using Twitter sentiment and weather. *2015 Systems and Information Engineering Design Symposium, SIEDS 2015*, (June 2015):63–68.

[45] Chen, Y., Ge, X., Yang, S., Hu, L., Li, J., and Zhang, J. (2023). A survey on multimodal knowledge graphs: Construction, completion and applications. *Mathematics*, 11(8).

[46] Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., and Duan, Z. (2020). Knowledge graph completion: A review. *IEEE Access*, 8:192435–192456.

[47] Cheng, B., Zhang, Y., and Shi, D. (2018). Ontology-based personalized learning path recommendation for course learning. In *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pages 531–535. IEEE.

[48] Chibeya, D., Wood, H., Cousins, S., Carter, K., Nyirenda, M. A., and Maseka, H. (2021). How do african elephants utilize the landscape during wet season? a habitat connectivity analysis for sioma ngwezi landscape in zambia. *Ecology and Evolution*, 11(21):14916–14931.

[49] Christian, E. (2005). Planning for the global earth observation system of systems (geoss). *Space Policy*, 21(2):105–109.

[50] Cleverly, J., Eamus, D., Edwards, W., Grant, M., Grundy, M. J., Held, A., Karan, M., Lowe, A. J., Prober, S. M., Sparrow, B., et al. (2019). Tern, australia's land observatory: addressing the global challenge of forecasting ecosystem responses to climate variability and change. *Environmental Research Letters*, 14(9):095004.

[51] Colborne, A. and Smit, M. (2020). Characterizing disinformation risk to open data in the post-truth era. *J. Data and Information Quality*, 12(3).

[52] Compton, M., Barnaghi, P., Bermudez, L., Garcia-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al. (2012). The ssn ontology of the w3c semantic sensor network incubator group. *Journal of Web Semantics*, 17:25–32.

[53] Contarinis, S., Pallikaris, A., and Nakos, B. (2020). The Value of Marine Spatial Open Data Infrastructures-Potentials of IHO S-100 Standard t Become the Universal Marine Data Model. *Journal of Marine Science and Engineering*, 8(8):564.

[54] Cooke, A., Smith, D., and Booth, A. (2012). Beyond pico: the spider tool for qualitative evidence synthesis. *Qualitative health research*, 22(10):1435–1443.

[55] Corcho, O., Chaves-Fraga, D., Toledo, J., Arenas-Guerrero, J., Badenes-Olmedo, C., Wang, M., Peng, H., Burrett, N., Mora, J., and Zhang, P. (2021). A high-level ontology network for ict infrastructures. In *The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20*, pages 446–462. Springer.

[56] Corlett, R. T. (2017). Frugivory and seed dispersal by vertebrates in tropical and subtropical asia: An update. *Global Ecology and Conservation*, 11:1–22.

[57] Correa, A. S., Zander, P.-O., and da Silva, F. S. C. (2018). Investigating open data portals automatically: A methodology and some illustrations. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, dg.o '18, New York, NY, USA. Association for Computing Machinery.

[58] Cota, G. et al. (2020). Best practices for implementing fair vocabularies and ontologies on the web. *Applications and practices in ontology design, extraction, and reasoning*, 49:39.

[59] Craglia, M., Hradec, J., Nativi, S., and Santoro, M. (2017). Exploring the depths of the global earth observation system of systems. *Big Earth Data*, 1(1-2):21–46.

[60] Critchlow, R., Plumptre, A. J., Driciru, M., Rwetsiba, A., Stokes, E. J., Tumwesigye, C., Wanyama, F., and Beale, C. M. (2015). Spatiotemporal trends of illegal activities from ranger-collected data in a Ugandan national park. *Conservation Biology*.

[61] Daniele, L., Hartog, F. d., and Roes, J. (2015). Created in close interaction with the industry: the smart appliances reference (saref) ontology. In *International Workshop Formal Ontologies Meet Industries*, pages 100–112. Springer.

[62] de Knegt, H. J., Eikelboom, J. A., van Langevelde, F., Spruyt, W. F., and Prins, H. H. (2021). Timely poacher detection and localization using sentinel animal movement. *Scientific reports*, 11(1):4596.

[63] Deepak, G., Rooban, S., and Santhanavijayan, A. (2021). A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. *Multimedia Tools and Applications*, 80(18):28061–28085.

[64] Department, S. W. (2020). Bornean elephant action plan for sabah 2020-2029.

[65] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., and Van de Walle, R. (2014). Rml: A generic language for integrated rdf mappings of heterogeneous data. *Ldow*, 1184.

[66] Dong, X. L. and Srivastava, D. (2013). Big data integration. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1245–1248.

[67] Dou, J., Qin, J., Jin, Z., and Li, Z. (2018). Knowledge graph based on domain ontology and natural language processing technology for chinese intangible cultural heritage. *Journal of Visual Languages & Computing*, 48:19–28.

[68] Du, N., Fathollahi-Fard, A. M., and Wong, K. Y. (2023). Wildlife resource conservation and utilization for achieving sustainable development in china: main barriers and problem identification. *Environmental Science and Pollution Research*, pages 1–20.

[69] Duan, W. and Chiang, Y.-Y. (2016). Building knowledge graph from public data for predictive analysis: a case study on predicting technology future in space and time. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 7–13.

[70] Duffy, R. (2022). Crime, security, and illegal wildlife trade: Political ecologies of international conservation. *Global Environmental Politics*, 22(2):23–44.

[71] Dunea, G. (2004). Privacy concerns. *BMJ*, 329(7464):519.

[72] Duporge, I. (2016). *Analysing the use of remote sensing & geospatial technology to combat wildlife crime in East and Southern Africa*.

[73] Dyo, V., Ellwood, S. A., Macdonald, D. W., Markham, A., Mascolo, C., Pásztor, B., Trigoni, N., and Wohlers, R. (2009). Wildlife and environmental monitoring using rfid and wsn technology. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, SenSys '09, page 371–372, New York, NY, USA. Association for Computing Machinery.

[74] Edemacu, K., Kim, J. W., Jang, B., and Park, H. K. (2019). Poacher detection in african game parks and reserves with iot: Machine learning approach. In *2019 International Conference on Green and Human Information Technology (ICGHIT)*, pages 12–17. IEEE.

[75] Ehrlinger, L. and Wöß, W. (2016a). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2.

[76] Ehrlinger, L. and Wöß, W. (2016b). Towards a definition of knowledge graphs. In *International Conference on Semantic Systems*.

[77] English, M., Gillespie, G., Ancrenaz, M., Ismail, S., Goossens, B., Nathan, S., and Linklater, W. (2014). Plant selection and avoidance by the bornean elephant (elephas maximus borneensis) in tropical forest: does plant recovery rate after herbivory influence food choices? *Journal of Tropical Ecology*, 30(4):371–379.

[78] Espinoza-Arias, P., Poveda-Villalón, M., and Corcho, O. (2020). Using lot methodology to develop a noise pollution ontology: a spanish use case. *Journal of Ambient Intelligence and Humanized Computing*, 11(11):4557–4568.

[79] et. al, E. (2018). Soil properties across primary forest, logged forest and oil palm plantation in sabah, malaysia.

[80] et al., W. (2020). Vegetation and habitat data for fragmented and continuous forest sites in sabah, malaysian borneo, 2017.

[81] Evans, L. J., Goossens, B., Davies, A. B., Reynolds, G., and Asner, G. P. (2020). Natural and anthropogenic drivers of bornean elephant movement strategies. *Global Ecology and Conservation*, 22:e00906.

[82] Evans, M., Yankov, D., Berkhin, P., Yudin, P., Teodorescu, F., and Wu, W. (2017). LiveMaps: Converting Map Images into Interactive Maps. SIGIR '17, pages 897–900. ACM.

[83] Fang, F., Nguyen, T. H., Pickles, R., Lam, W. Y., Clements, G. R., An, B., Singh, A., Tambe, M., and Lemieux, A. (2016). Deploying PAWS : Field Optimization of the Protection Assistant for Wildlife Security. *Proceedings of the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference*.

[84] Fang, F., Nguyen, T. H., Sinha, A., Gholami, S., Plumptre, A., Joppa, L., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Critchlow, R., and Beale, C. M. (2017). Predicting poaching for wildlife Protection. *IBM Journal of Research and Development*, 61(6).

[85] Feng, Z.-Y., Wu, X.-H., Ma, J.-L., Li, M., He, G.-F., Cao, D.-S., and Yang, G.-P. (2023). Dkade: a novel framework based on deep learning and knowledge graph for identifying adverse drug events and related medications. *Briefings in Bioinformatics*, page bbad228.

[86] Ferber, A., Griffin, E., Dilkina, B., Keskin, B., and Gore, M. (2023). Predicting wildlife trafficking routes with differentiable shortest paths. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 460–476. Springer.

[87] Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering.

[88] Foundation, S. (2010). Ten Principles for Opening Up Government Information. *Sunlight Foundation*, (October 2007):3.

[89] Frey, J. and Hellmann, S. (2021). Fair linked data - towards a linked data backbone for users and machines. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 431–435, New York, NY, USA. Association for Computing Machinery.

[90] Frey, R. M., Hardjono, T., Smith, C., Erhardt, K., and Pentland, A. S. (2017). Secure sharing of geospatial wildlife data. In *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data*, GeoRich '17, New York, NY, USA. Association for Computing Machinery.

[91] Frummet, A., Elsweiler, D., and Ludwig, B. (2022). "what can i cook with these ingredients?"-understanding cooking-related information needs in conversational search. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–32.

[92] Gangemi, A. and Presutti, V. (2009). *Ontology Design Patterns*, pages 221–243. Springer Berlin Heidelberg, Berlin, Heidelberg.

[93] Garai, M. E., Roos, T., Eggeling, T., Ganswindt, A., Pretorius, Y., and Henley, M. (2022). Developing welfare parameters for african elephants (loxodonta africana) in fenced reserves in south africa. *Plos one*, 17(3):e0264931.

[94] Garijo, D. (2017). Widoco: a wizard for documenting ontologies. In *International Semantic Web Conference*, pages 94–102. Springer.

[95] Garijo, D., Corcho, O., and Poveda-Villalón, M. (2021). Foops!: An ontology pitfall scanner for the fair principles. 2980.

[96] Gharaibeh, A., Salahuddin, M. A., Hussini, S. J., Khreishah, A., Khalil, I., Guizani, M., and Al-Fuqaha, A. (2017). Smart cities: A survey on data management, security, and enabling technologies. *IEEE Communications Surveys Tutorials*, 19(4):2456–2501.

[97] Gholami, S., Ford, B., Fang, F., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., and Mabonga, J. (2017). Taking It for a Test Drive: A Hybrid Spatio-Temporal Model for Wildlife Poaching Prediction Evaluated Through a Controlled Field Test. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10536 LNAI(1):292–304.

[98] Gholami, S., McCarthy, S., Dilkina, B., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., Mabonga, J., Okello, T., and Enyel, E. (2018). Adversary models account for imperfect crime data: Forecasting and planning against real-world poachers. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2:823–831.

[99] Gilbert, N. A., Blommel, C. M., Farr, M. T., Green, D. S., Holekamp, K. E., and Zipkin, E. F. (2024). A multispecies hierarchical model to integrate count and distance-sampling data. *Ecology*, page e4326.

[100] Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2006). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media.

[101] Gómez-Pérez, A. and Suárez-Figueroa, M. C. (2009). Neon methodology for building ontology networks: a scenario-based methodology.

[102] Gómez-Pérez, A. and Suárez-Figueroa, M. C. (2009). Scenarios for building ontology networks within the NeOn Methodology. *K-CAP'09 - Proceedings of the 5th International Conference on Knowledge Capture*, pages 183–184.

[103] Gómez-pérez, A. A., Motta, E., and Suárez-figueroa, M. C. (2005). Introduction to the NeOn Methodology Introduction to the NeOn Methodology. *Cycle*, (c).

[104] Gonzalez, L. F., Montes, G. A., Puig, E., Johnson, S., Mengersen, K., and Gaston, K. J. (2016). Unmanned aerial vehicles (uavs) and artificial intelligence revolutionizing wildlife monitoring and conservation. *Sensors*, 16(1):97.

[105] Google (Accessed March 31, 2023). Google Scholar. https://scholar.google.com.

[106] Goossens, B., Sharma, R., Othman, N., Kun-Rodrigues, C., Sakong, R., Ancrenaz, M., Ambu, L. N., Jue, N. K., O'Neill, R. J., Bruford, M. W., et al. (2016). Habitat fragmentation and genetic diversity in natural populations of the bornean elephant: Implications for conservation. *Biological Conservation*, 196:80–92.

[107] Gordon, S. N., Murphy, P. J., Gallo, J. A., Huber, P., Hollander, A., Edwards, A., and Jankowski, P. (2021). People, projects, organizations, and products: Designing a knowledge graph to support multi-stakeholder environmental planning and design. *ISPRS International Journal of Geo-Information*, 10(12):823.

[108] Gore, M. L., Griffin, E., Dilkina, B., Ferber, A., Griffis, S. E., Keskin, B. B., and Macdonald, J. (2023). Advancing interdisciplinary science for disrupting wildlife trafficking networks. *Proceedings of the National Academy of Sciences*, 120(10):e2208268120.

[109] Grewal, R. (2009). *The book of Ganesha*. Penguin Books India.

[110] Gries, C., Hanson, P. C., O'Brien, M., Servilla, M., Vanderbilt, K., and Waide, R. (2023). The environmental data initiative: Connecting the past to the future through data reuse. *Ecology and Evolution*, 13(1):e9592.

[111] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

[112] Grüninger, M. and Fox, M. S. (1995). Methodology for the design and evaluation of ontologies.

[113] Gurumurthy, S., Yu, L., Zhang, C., Jin, Y., Li, W., Zhang, X., and Fang, F. (2018). Exploiting data and human knowledge for predicting wildlife poaching. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–8.

[114] Gutiérrez, C. and Sequeda, J. F. (2021). Knowledge graphs. *Communications of the ACM*, 64(3):96–104.

[115] Haas, T. C. and Ferreira, S. M. (2015). Federated databases and actionable intelligence: using social network analysis to disrupt transnational wildlife trafficking criminal networks. *Security Informatics*, 4(1):1–14.

[116] Haas, T. C. and Ferreira, S. M. (2018). Finding politically feasible conservation policies: the case of wildlife trafficking. *Ecological Applications*, 28(2):473–494.

[117] Hegde, K. and Sen, S. (2022). Iot-based anti-poaching technology to save wildlife. In *Proceedings of International Conference on Advanced Computing Applications: ICACA 2021*, pages 41–52. Springer.

[118] Heist, N. and Paulheim, H. (2019). Uncovering the Semantics of Wikipedia Categories. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11778 LNCS:219–236.

[119] Herbert, C., Haya, B. K., Stephens, S. L., and Butsic, V. (2022). Managing nature-based solutions in fire-prone ecosystems: Competing management objectives in california forests evaluated at a landscape scale. *Frontiers in Forests and Global Change*, 5:957189.

[120] Heyvaert, P., De Meester, B., Dimou, A., and Verborgh, R. (2018). Declarative Rules for Linked Data Generation at your Fingertips! In *Proceedings of the 15th ESWC: Posters and Demos*.

[121] Hofer, H., Campbell, K. L., East, M. L., and Huish, S. A. (2000). Modeling the spatial distribution of the economic costs and benefits of illegal game meat hunting in the serengeti. *Natural Resource Modeling*.

[122] Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge graphs. *ACM Comput. Surv.*, 54(4).

[123] Houe, H., Nielsen, S. S., Nielsen, L. R., Ethelberg, S., and Mølbak, K. (2019). Opportunities for improved disease surveillance and control by use of integrated data on animal and human health. *Frontiers in Veterinary Science*, 6:301.

[124] Hu, R., Yan, Z., Ding, W., and Yang, L. T. (2020). A survey on data provenance in IoT. *World Wide Web*, 23(2):1441–1463.

[125] Huang, L., Yu, C., Chi, Y., Qi, X., and Xu, H. (2019). Towards smart healthcare management based on knowledge graph technology. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, pages 330–337.

[126] Hunter, E. A., Blake, S., Cayot, L. J., and Gibbs, J. P. (2021). Role in ecosystems. In *Galapagos giant tortoises*, pages 299–315. Elsevier.

[127] Iannacone, M., Bohn, S., Nakamura, G., Gerth, J., Huffer, K., Bridges, R., Ferragut, E., and Goodall, J. (2015). Developing an ontology for cyber security knowledge graphs. In *Proceedings of the 10th Annual Cyber and Information Security Research Conference*, pages 1–4.

[128] Ihwagi, F. W., Skidmore, A. K., Wang, T., Bastille-Rousseau, G., Toxopeus, A. G., and Douglas-Hamilton, I. (2019). Poaching lowers elephant path tortuosity: implications for conservation. *The Journal of Wildlife Management*, 83(5):1022–1031.

[129] Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., and Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, 6(3):4219–4225.

[130] Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., et al. (2020). Fair principles: interpretations and implementation considerations.

[131] James, A., Kanyamibwa, S., Green, M. J., and Anderson, T. (2001). Sustainable financing for protected areas in sub-saharan africa and the caribbean. *The politics and economics of park management*, pages 69–87.

[132] Janowicz, K., Haller, A., Cox, S. J., Le Phuoc, D., and Lefrançois, M. (2019). SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 56:1–10.

[133] Jin, Y., Liu, Y., and Wang, J. (2024). Reducing poaching in elephant populations on random forest algorithm. *Highlights in Business, Economics and Management*, 30:395–400.

[134] Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Yeumo, E. D., Emonet, V., Graybeal, J., Laporte, M.-A., Musen, M. A., Pesce, V., et al. (2018). Agroportal: A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144:126–143.

[135] Kamminga, J., Ayele, E., Meratnia, N., and Havinga, P. (2018). Poaching detection technologies—a survey. *Sensors*, 18(5):1474.

[136] Kamran, A. B., Abro, B., and Basharat, A. (2023). Semantichadith: An ontology-driven knowledge graph for the hadith corpus. *Journal of Web Semantics*, page 100797.

[137] Kang, H.-W. and Kang, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PloS one*, 12(4):e0176244.

[138] Kar, D., Ford, B., Gholami, S., Fang, F., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., et al. (2017a). Cloudy with a chance of poaching: Adversary behavior modeling and forecasting with real-world poaching data. International Conference on Autonomous Agents and Multiagent Systems.

[139] Kar, D., Ford, B., Gholami, S., Fang, F., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., and Mabonga, J. (2017b). Cloudy with a chance of poaching: Adversary behavior modeling and forecasting with real-world poaching data. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*.

[140] Keet, C. M. (2020). The african wildlife ontology tutorial ontologies. *Journal of Biomedical Semantics*, 11(1).

[141] Kejriwal, M. (2019). *Domain-specific knowledge graph construction*. Springer.

[142] Kerley, G. I. and Landman, M. (2006). The impacts of elephants on biodiversity in the eastern cape subtropical thickets: elephant conservation. *South African Journal of Science*, 102(9):395–402.

[143] Keskin, B. B., Griffin, E. C., Prell, J. O., Dilkina, B., Ferber, A., MacDonald, J., Hilend, R., Griffis, S., and Gore, M. L. (2023). Quantitative investigation of wildlife trafficking supply chains: A review. *Omega*, 115:102780.

[144] Kitchin, R. and Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, 80:463–475.

[145] Kiv, D., Allabadi, G., Kaplan, B., and Kravets, R. (2022). Smol: Sensing soil moisture using lora. In *Proceedings of the 1st ACM Workshop on No Power and Low Power Internet-of-Things*, LP-IoT'21, page 21–27, New York, NY, USA. Association for Computing Machinery.

[146] Komninos, N., Bratsas, C., Kakderi, C., and Tsarchopoulos, P. (2019). Smart city ontologies: Improving the effectiveness of smart city applications. *Journal of Smart Cities*, 1(1):31–46.

[147] Kovács, K. Z., Hemment, D., Woods, M., van der VELDEN, N. K., Xaver, A., Gi Esen, R. H., Burton, V. J., Garrett, N. L., Zappa, L., Long, D., Dobos, E., and Skalsky, R. (2019). Citizen observatory based soil moisture monitoring – The GROW example. *Hungarian Geographical Bulletin*, 68(2):119–139.

[148] Krötzsch, M. (2017). Ontologies for knowledge graphs? In *Description Logics*.

[149] Kucera, J., Chlapek, D., Klímek, J., and Necaskỳ, M. (2015). Methodologies and best practices for open data publication. In *DATESO*, pages 52–64.

[150] Kumar, P. R., Wan, A. T., and Suhaili, W. S. H. (2020). Exploring Data Security and Privacy Issues in Internet of Things Based on Five-Layer Architecture. *International journal of communication networks and information security*, 12(1):108–121.

[151] Lämmel, P., Dittwald, B., Bruns, L., Tcholtchev, N., Glikman, Y., Cuno, S., Flügge, M., and Schieferdecker, I. (2020). Metadata harvesting and quality assurance within open urban platforms. *J. Data and Information Quality*, 12(4).

[152] Lane, M. A. and Edwards, J. L. (2007). The global biodiversity information facility (gbif). *Systematics Association special volume*, 73:1.

[153] Lavadinović, V. M., Islas, C. A., Chatakonda, M. K., Marković, N., and Mbiba, M. (2021). Mapping the research landscape on poaching: A decadal systematic review. *Frontiers in Ecology and Evolution*, 9:630990.

[154] Lee, E. (2011). Reflections on the decadal-scale response of coastal cliffs to sea-level rise. *Quarterly Journal of Engineering Geology and Hydrogeology*, 44(4):481–489.

[155] Li, T., Gao, C., Jiang, L., Pedrycz, W., and Shen, J. (2019). Publicly verifiable privacy-preserving aggregation and its application in IoT. *Journal of Network and Computer Applications*, 126(October 2018):39–44.

[156] Li, Z., Li, H., and Meng, L. (2023). Model compression for deep neural networks: A survey. *Computers*, 12(3).

[157] Ling, L. E., Ariffin, M., and Abd Manaf, L. (2016). A qualitative analysis of the main threats to asian elephant conservation. *Gajah*, 44:16–22.

[158] Liu, X., Yang, T., and Yan, B. (2015). Internet of things for wildlife monitoring. In *2015 IEEE/CIC International Conference on Communications in China-Workshops (CIC/ICCC)*, pages 62–66. IEEE.

[159] Liu, Y., Qiu, M., Liu, C., and Guo, Z. (2017). Big data challenges in ocean observation: a survey. *Personal and Ubiquitous Computing*, 21:55–65.

[160] Liu, Y. N., Wang, Y. P., Wang, X. F., Xia, Z., and Xu, J. F. (2019). Privacy-preserving raw data collection without a trusted authority for IoT. *Computer Networks*, 148:340–348.

[161] Lnenicka, M. and Nikiforova, A. (2021). Transparency-by-design: What is the role of open data portals? *Telematics and Informatics*, 61:101605.

[162] López, M. F., Gómez-Pérez, A., Sierra, J. P., and Sierra, A. P. (1999). Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems*, 14(1):37–46.

[163] Lynn, M. S. and Jumail, A. (2021). The danau girang field centre: Field station profile. *ECOTROPICA*, 23(1/2):202103–202103.

[164] Ma, M., Preum, S. M., Ahmed, M. Y., Tärneberg, W., Hendawi, A., and Stankovic, J. A. (2019). Data sets, modeling, and decision making in smart cities: A survey. *ACM Transactions on Cyber-Physical Systems*, 4(2).

[165] Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279–296. Meta-information systems and ontologies. A Special Feature from the 5th International Conference on Ecological Informatics ISEI5, Santa Barbara, CA, Dec. 4–7, 2006.

[166] Mane, V., Nikude, P., Patil, T., and Tambe, P. (2024). Wildlife classification using convolutional neural networks (cnn). In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 1046–1053. IEEE.

[167] Mansour, E., Srinivas, K., and Hose, K. (2022). Federated data science to break down silos [vision]. *ACM SIGMOD Record*, 50(4):16–22.

[168] Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., and Yaqoob, I. (2017). Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. *IEEE access*, 5:5247–5261.

[169] Mason, T. and Dhoop, T. (2017). Cover photograph: Datawell Directional Waverider Mk III in Weymouth Bay Photo courtesy of Fugro GB Marine Limited National Network of Regional Coastal Monitoring Programmes of England Quality Assurance & Quality Control of Wave Data.

[170] Mäyrä, J., Keski-Saari, S., Kivinen, S., Tanhuanpää, T., Hurskainen, P., Kullberg, P., Poikolainen, L., Viinikka, A., Tuominen, S., Kumpula, T., et al. (2021). Tree species classification from airborne hyperspectral and lidar data using 3d convolutional neural networks. *Remote Sensing of Environment*, 256:112322.

[171] McCarthy, D. P., Donald, P. F., Scharlemann, J. P., Buchanan, G. M., Balmford, A., Green, J. M., Bennun, L. A., Burgess, N. D., Fishpool, L. D., Garnett, S. T., et al. (2012). Financial costs of meeting global biodiversity conservation targets: current spending and unmet needs. *Science*, 338(6109):946–949.

[172] McDaniel, M. and Storey, V. (2019a). Evaluating domain ontologies: Clarification, classification, and challenges. *ACM computing surveys*, 52(4):1–44.

[173] McDaniel, M. and Storey, V. C. (2019b). Evaluating domain ontologies: clarification, classification, and challenges. *ACM Computing Surveys (CSUR)*, 52(4):1–44.

[174] Mendoza, A. P., Shanee, S., Cavero, N., Lujan-Vega, C., Ibanez, Y., Rynaby, C., Villena, M., Murillo, Y., Olson, S. H., Perez, A., et al. (2022). Domestic networks contribute to the diversity and composition of live wildlife trafficked in urban markets in peru. *Global Ecology and Conservation*, 37:e02161.

[175] Michel, F., Djimenou, L., Zucker, C. F., and Montagnat, J. (2017). *xR2RML: Relational and non-relational databases to RDF mapping language*. PhD thesis, CNRS.

[176] Miller, H., Clifton, K., Akar, G., Tufte, K., Gopalakrishnan, S., MacArthur, J., Irwin, E., Ramnath, R., and Stiles, J. (2021). Urban Sustainability Observatories: Leveraging Urban Experimentation for Sustainability Science and Policy. *Harvard Data Science Review*, 3(2). https://hdsr.mitpress.mit.edu/pub/zunejoo2.

[177] Mireku Kwakye, M. (2019). Semantic data warehouse modelling for trajectories. *arXiv e-prints*, pages arXiv–1904.

[178] Moustaka, V., Vakali, A., and Anthopoulos, L. G. (2018). A systematic review for smart city data analytics. *ACM Comput. Surv.*, 51(5).

[179] Moustard, F., Haklay, M., Lewis, J., Albert, A., Moreu, M., Chiaravalloti, R., Hoyte, S., Skarlatidou, A., Vittoria, A., Comandulli, C., et al. (2021). Using sapelli in the field: methods and data for an inclusive citizen science. *Frontiers in Ecology and Evolution*, 9:638870.

[180] Mukwazvure, A. and Magadza, T. (2014). A survey on anti-poaching strategies. *International Journal of Science and Research*, 3(6):1064–1066.

[181] Naderifar, M., Goli, H., and Ghaljaie, F. (2017). Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in development of medical education*, 14(3).

[182] Naidoo, R., Fisher, B., Manica, A., and Balmford, A. (2016). Estimating economic losses to tourism in africa from the illegal killing of elephants. *Nature communications*, 7(1):13379.

[183] Nathan, R., Monk, C. T., Arlinghaus, R., Adam, T., Alós, J., Assaf, M., Baktoft, H., Beardsworth, C. E., Bertram, M. G., Bijleveld, A. I., et al. (2022). Big-data approaches lead to an increased understanding of the ecology of animal movement. *Science*, 375(6582):eabg1780.

[184] Neil, E., Madsen, J. K., Carrella, E., Payette, N., and Bailey, R. (2020). Agent-based modelling as a tool for elephant poaching mitigation. *Ecological modelling*, 427:109054.

[185] Nguyen, T. H., Sinha, A., Gholami, S., Plumptre, A., Joppa, L., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Critchlow, R., and Beale, C. M. (2016). CAPTURE: A new predictive anti-poaching tool for wildlife protection. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, pages 767–775.

[186] Ning, Y., Liu, H., Wang, H., Zeng, Z., and Xiong, H. (2023). Uukg: Unified urban knowledge graph dataset for urban spatiotemporal prediction. *arXiv preprint arXiv:2306.11443*.

[187] Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725.

[188] Noura, M., Atiquzzaman, M., and Gaedke, M. (2019). Interoperability in internet of things: Taxonomies and open challenges. *Mobile Networks and Applications*, 24:796–809.

[189] Novo, O. and Francesco, M. D. (2020). Semantic interoperability in the iot: Extending the web of things architecture. *ACM Trans. Internet Things*, 1(1).

[190] Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology.

[191] Nuwer, R. L. (2018). *Poached: inside the dark world of wildlife trafficking*. Hachette UK.

[192] of Standards, N. I. and Technology (2010). QUDT - quantities, units, dimensions and data types ontology. http://qudt.org/.

[193] Oklander, L., Ang, A., and Ikemeh, R. A. (2024). Advancing conservation of threatened primates. *Oryx*, 58(2):137–138.

[194] Oliver, J. L., Brereton, M., Watson, D. M., and Roe, P. (2019). Listening to save wildlife: Lessons learnt from use of acoustic technology by a species recovery team. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, page 1335–1348, New York, NY, USA. Association for Computing Machinery.

[195] Open Knowledge Foundation (2021). Open data handbook. https://opendatahandbook.org/. Accessed: 2024-03-07.

[196] Page, R. D. (2019). Ozymandias: a biodiversity knowledge graph. *PeerJ*, 7:e6739.

[197] Pahuja, V., Wang, B., Latapie, H., Srinivasa, J., and Su, Y. (2023). A retrieve-and-read framework for knowledge graph link prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1992–2002.

[198] Paranayapa, T., Ranasinghe, P., Ranmal, D., Meedeniya, D., and Perera, C. (2024). A comparative study of preprocessing and model compression techniques in deep learning for forest sound classification. *Sensors*, 24(4).

[199] Park, N., Serra, E., Snitch, T., and Subrahmanian, V. (2015). Ape: A data-driven, behavioral model-based anti-poaching engine. *IEEE Transactions on Computational Social Systems*, 2(2):15–37.

[200] Parr, C. S., Wilson, M. N., Leary, M. P., Schulz, K. S., Lans, M. K., Walley, M. L., Hammock, J. A., Goddard, M. A., Rice, M. J., Studer, M. M., et al. (2014). The encyclopedia of life v2: providing global access to knowledge about life on earth. *Biodiversity data journal*, (2).

[201] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

[202] Pearce, H. (2020). The (UK) Freedom of Information Act's disclosure process is broken: where do we go from here? *Information and Communications Technology Law*, 29(3):354–390.

[203] Pendleton, M., Garcia-Lebron, R., Cho, J.-H., and Xu, S. (2017). A survey on systems security metrics. *ACM computing surveys*, 49(4):1–35.

[204] Penev, L., Dimitrova, M., Senderov, V., Zhelezov, G., Georgiev, T., Stoev, P., and Simov, K. (2019). Openbiodiv: a knowledge graph for literature-extracted linked open data in biodiversity science. 7(2):38.

[205] Perera, B. (2009). The human-elephant conflict: A review of current status and mitigation methods. *Gajah*, 30(2009):41–52.

[206] Perera, C., Qin, Y., Estrella, J. C., Reiff-Marganiec, S., and Vasilakos, A. V. (2017). Fog computing for sustainable smart cities: A survey. *ACM Computing Surveys*, 50(3).

[207] Perez-Castillo, R., Carretero, A. G., Rodriguez, M., Caballero, I., Piattini, M., Mate, A., Kim, S., and Lee, D. (2018). Data Quality Best Practices in IoT Environments. In *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, pages 272–275.

[208] Peroni, S., Shotton, D., and Vitali, F. (2012). The live owl documentation environment: a tool for the automatic generation of ontology documentation. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 398–412. Springer.

[209] Picco, G. P., Molteni, D., Murphy, A. L., Ossi, F., Cagnacci, F., Corrà, M., and Nicoloso, S. (2015). Geo-referenced proximity detection of wildlife with wildscope: Design and characterization. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, IPSN '15, page 238–249, New York, NY, USA. Association for Computing Machinery.

[210] Pinto, H. S., Staab, S., and Tempich, C. (2004). Diligent: Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In *ECAI*.

[211] Poveda-Villalón, M., Fernández-Izquierdo, A., Fernández-López, M., and García-Castro, R. (2022). Lot: An industrial oriented ontology engineering framework. *Engineering Applications of Artificial Intelligence*, 111.

[212] Price, B., Briscoe, A., Misra, R., and Broad, G. (2020). Evaluation of dna barcode libraries used in the uk and developing an action plan to fill priority gaps. Technical report, Natural England.

[213] Raad, J. and Cruz, C. (2015). A survey on ontology evaluation methods. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

[214] Raffik, R., Swetha, M. M., Sathya, R. R., Vaishali, V., Adithya, B. M., and Balavedhaa, S. (2024). Integration of unmanned aerial vehicle systems with machine learning algorithms for wildlife monitoring and conservation. In *Applications of Machine Learning in UAV Networks*, pages 121–159. IGI Global.

[215] Raimond, Y., Scott, T., Oliver, S., Sinclair, P., and Smethurst, M. (2010). *Use of Semantic Web technologies on the BBC Web Sites*, pages 263–283. Springer US, Boston, MA.

[216] Ramapriyan, H. K. and Moses, J. (2010). Nasa's earth science data systems: Lessons learned and future directions. In *Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop*, pages 1–9.

[217] Raskin, R. G. and Pan, M. J. (2005). Knowledge representation in the semantic web for earth and environmental terminology (sweet). *Computers Geosciences*, 31(9):1119–1125. Application of XML in the Geosciences.

[218] Revathi, T. S., Ramaraj, N., and Chithra, S. (2020). Tracy-Singh Product and Genetic Whale Optimization Algorithm for Retrievable Data Perturbation for Privacy Preserved Data Publishing in Cloud Computing. *Computer Journal*, 63(2):239–253.

[219] Ribeiro, J., Bingre, P., Strubbe, D., Santana, J., Capinha, C., Araújo, M. B., and Reino, L. (2022). Exploring the effects of geopolitical shifts on global wildlife trade. *Bioscience*, 72(6):560–572.

[220] Ritts, M. and Bakker, K. (2021). Conservation acoustics: Animal sounds, audible natures, cheap nature. *Geoforum*, 124:144–155.

[221] Rivera, J. W. (2023). Making the best of bad data: A longitudinal analysis of elephant poaching in africa from 2002 to 2018. *Journal of Economic Criminology*, 2:100020.

[222] Rüegg, J., Gries, C., Bond-Lamberty, B., Bowen, G. J., Felzer, B. S., McIntyre, N. E., Soranno, P. A., Vanderbilt, K. L., and Weathers, K. C. (2014). Completing the data life cycle: using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment*, 12(1):24–30.

[223] Ryen, V., Soylu, A., and Roman, D. (2022). Building semantic knowledge graphs from (semi-)structured data: A review. *Future Internet*, 14(5).

[224] S, V., G, T., Nandi, S., M, S., and P, A. (2021). Forest fire detection and guiding animals to a safe area by using sensor networks and sound. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pages 473–476.

[225] Santipantakis, G. M., Vouros, G. A., Doulkeridis, C., Vlachou, A., Andrienko, G., Andrienko, N., Fuchs, G., Garcia, J. M. C., and Martinez, M. G. (2017). Specification of semantic trajectories supporting data transformations for analytics: The datacron ontology. In *Proceedings of the 13th International Conference on Semantic Systems*, Semantics2017, page 17–24, New York, NY, USA. Association for Computing Machinery.

[226] Scholl, V. M., Cattau, M. E., Joseph, M. B., and Balch, J. K. (2020). Integrating national ecological observatory network (neon) airborne remote sensing and in-situ data for optimal tree species classification. *Remote Sensing*, 12(9).

[227] Sikos, L. F. (2023). Cybersecurity knowledge graphs. *Knowledge and Information Systems*, 65(9):3511–3531.

[228] Singh, A. and Anand, P. (2013). State of art in ontology development tools. *International Journal*, 2(7).

[229] Singh, R., Gan, M., Barlow, C., Long, B., Mcvey, D., De Kock, R., Gajardo, O. B., Avino, F. S., and Belecky, M. (2020). What do rangers feel? perceptions from asia, africa and latin america. *Parks*, 26(1):63–76.

[230] Siow, E., Tiropanis, T., and Hall, W. (2018). Analytics for the internet of things: A survey. *ACM Computing Surveys*, 51(4).

[231] Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, 5(2):51–53.

[232] Smith, B. R., Root-Gutteridge, H., Butkiewicz, H., Dassow, A., Fontaine, A. C., Markham, A., Owens, J., Schindler, L., Wijers, M., and Kershenbaum, A. (2021). Acoustic localisation of wildlife with low-cost equipment: lower sensitivity, but no loss of precision. *Wildlife Research*.

[233] Smith, L. and Turner, M. (2019). Building the Urban Observatory: Engineering the largest set of publicly available real-time environmental urban data in the UK. *Geophysical Research Abstracts*, 21:1.

[234] Srivastava, A. (2018). *Mastering Kibana 6. x: Visualize Your Elastic Stack Data with Histograms, Maps, Charts, and Graphs*. Packt Publishing, Limited, Birmingham.

[235] Stall, S., Martone, M. E., Chandramouliswaran, I., Crosas, M., Federer, L., Gautier, J., Hahnel, M., Larkin, J., Lowenberg, D., Pfeiffer, N., Sim, I., Smith, T., Van Gulick, A. E., Walker, E., Wood, J., Zaringhalam, M., and Zigoni, A. (2020). Generalist repository comparison chart. Thank you the American Geophysical Union for designing the document.

[236] Stehle, S. and Kitchin, R. (2020). Real-time and archival data visualisation techniques in city dashboards. *International Journal of Geographical Information Science*, 34(2):344–366.

[237] Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997). Data Quality in Context. *Commun. ACM*, 40(5):103–110.

[238] Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2012). The neon methodology for ontology engineering. In *Ontology engineering in a networked world*, pages 9–34. Springer.

[239] Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2015). The NeOn Methodology framework: Ascenario-based methodology for ontologydevelopment. *Applied Ontology*, 10(2):107–145.

[240] Suárez-Figueroa, M. C., Gómez-Pérez, A., and Villazón-Terrazas, B. (2009). How to write and use the ontology requirements specification document. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 966–982. Springer.

[241] Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., et al. (2014). The ebird enterprise: An integrated approach to development and application of citizen science. *Biological conservation*, 169:31–40.

[242] Sure, Y., Staab, S., and Studer, R. (2004). *On-To-Knowledge Methodology (OTKM)*, pages 117–132. Springer Berlin Heidelberg, Berlin, Heidelberg.

[243] Sutter, R. D., Wainscott, S. B., Boetsch, J. R., Palmer, C. J., and Rugg, D. J. (2015). Practical guidance for integrating data management into long-term ecological monitoring projects. *Wildlife Society Bulletin*, 39(3):451–463.

[244] Taleb, I., Serhani, M. A., and Dssouli, R. (2018). Big Data Quality: A Survey. In *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 166–173.

[245] Tianxing, M., Lushnov, M., Ignatov, D. I., Shichkina, Y. A., Zhukova, N. A., and Vodyaho, A. I. (2021). An ontology-based approach to the analysis of the acid-base state of patients at operative measures. *PeerJ Computer Science*, 7.

[246] Tompson, L., Johnson, S., Ashby, M., Perkins, C., and Edwards, P. (2015). Uk open source crime data: accuracy and possibilities for research. *Cartography and geographic information science*, 42(2):97–111.

[247] Urbano, F., Viterbi, R., Pedrotti, L., Vettorazzo, E., Movalli, C., and Corlatti, L. (2024). Enhancing biodiversity conservation and monitoring in protected areas through efficient data management. *Environmental Monitoring and Assessment*, 196(1):12.

[248] UWA, W. et al. (2012). Protecting the wildlife corridors of the queen elizabeth conservation area.

[249] Van Assche, D., Delva, T., Haesendonck, G., Heyvaert, P., De Meester, B., and Dimou, A. (2023). Declarative rdf graph generation from heterogeneous (semi-) structured data: A systematic literature review. *Journal of Web Semantics*, 75:100753.

[250] Van Assche, D., Delva, T., Haesendonck, G., Heyvaert, P., De Meester, B., and Dimou, A. (2023). Declarative rdf graph generation from heterogeneous (semi-)structured data: A systematic literature review. *Journal of Web Semantics*, 75:100753.

[251] Van Assche, D., Delva, T., Heyvaert, P., De Meester, B., and Dimou, A. (2021). Towards a more human-friendly knowledge graph generation and publication. In *International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks*.

[252] van Uhm, D. (2023). The criminalization of the trade in wildlife. In *Organized Crime in the 21st Century: Motivations, Opportunities, and Constraints*, pages 155–169. Springer.

[253] Varghese, A. O., Suryavanshi, A. S., and Jha, C. S. (2022). *Geospatial Applications in Wildlife Conservation and Management*, pages 727–750. Springer International Publishing, Cham.

[254] Vihervaara, P., Anttila, S., Kullberg, P., Härmä, P., Törmä, M., Jussila, T., Aapala, K., Heikkinen, R., Mäyrä, J., Kervinen, M., et al. (2021). Finnish ecosystem observatory (feo)-operationalizing remote sensing analyses for threatened habitats and biodiversity monitoring. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 735–738. IEEE.

[255] Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

[256] W3C Semantic Sensor Networks Working Group (2019). Semantic sensor network ontology (ssn) version 2. https://www.w3.org/TR/vocab-ssn/. [Online; accessed March 24, 2023].

[257] Waldron, A., Mooers, A. O., Miller, D. C., Nibbelink, N., Redding, D., Kuhn, T. S., Roberts, J. T., and Gittleman, J. L. (2013). Targeting global conservation funding to limit immediate biodiversity declines. *Proceedings of the National Academy of Sciences*, 110(29):12144–12148.

[258] Wang, C., Lin, Z., Yang, X., Sun, J., Yue, M., and Shahabi, C. (2022a). Hagen: Homophily-aware graph convolutional recurrent network for crime forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4193–4200.

[259] Wang, P., Fu, L., Patton, E. W., McGuinness, D. L., Dein, F. J., and Bristol, R. S. (2012). Towards semantically-enabled exploration and analysis of environmental ecosystems. In *2012 IEEE 8th International Conference on E-Science*, pages 1–8.

[260] Wang, S., Du, Z., Ding, M., Rodriguez-Paton, A., and Song, T. (2022b). Kg-dti: a knowledge graph based deep learning method for drug-target interaction predictions and alzheimer's disease drug repositions. *Applied Intelligence*, 52(1):846–857.

[261] Wang, V. and Shepherd, D. (2020). Exploring the extent of openness of open government data – A critique of open government datasets in the UK. *Government Information Quarterly*, 37(1):101405.

[262] Wannous, R., Malki, J., Bouju, A., and Vincent, C. (2017). Trajectory ontology inference considering domain and temporal dimensions—application to marine mammals. *Future Generation Computer Systems*, 68:491–499.

[263] Watch, G. F. (2002). Global forest watch. *World Resources Institute, Washington, DC Available from http://www. globalforestwatch. org (accessed March 2002)*.

[264] Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., and Dwivedi, Y. K. (2017). Open data and its usability: an empirical view from the Citizen's perspective. *Information Systems Frontiers*, 19(2):285–300.

[265] Werder, K., Ramesh, B., and Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, 13(2):1–23.

[266] Wieder, P. and Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in Big Data*, 5.

[267] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

[268] Woods, M., Hemment, D., Ajates, R., Cobley, A., Xaver, A., and Konstantakopoulos, G. (2020). GROW Citizens' Observatory: Leveraging the power of citizens, open data and technology to generate engagement, and action on soil policy and soil moisture monitoring. *IOP Conference Series: Earth and Environmental Science*, 509(1):10–12.

[269] Xie, R., Xu, Y., Ma, M., and Wang, Z. (2024). Fish physiologically based toxicokinetic modeling approach for in vitro–in vivo and cross-species extrapolation of endocrine-disrupting chemicals in risk assessment. *Environmental Science & Technology*.

[270] Xu, L., Gholami, S., McCarthy, S., Dilkina, B., Plumptre, A., Tambe, M., Singh, R., Nsubuga, M., Mabonga, J., Driciru, M., Wanyama, F., Rwetsiba, A., Okello, T., and Enyel, E. (2020). Stay ahead of poachers: Illegal wildlife poaching prediction and patrol planning under uncertainty with field test evaluations (short version). In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1898–1901.

[271] Yan, W., Shi, Y., Ji, Z., Sui, Y., Tian, Z., Wang, W., and Cao, Q. (2023). Intelligent predictive maintenance of hydraulic systems based on virtual knowledge graph. *Engineering Applications of Artificial Intelligence*, 126:106798.

[272] Yang, C., Wang, H., Sai, J., Zhang, P., and Yan, M. (2020). Construction of knowledge graph of forest musk deer based on bilstm-crf model and dpa method. In *IOP Conference Series: Earth and Environmental Science*, volume 615, page 012008. IOP Publishing.

[273] Yang, H.-F., Chen, C.-H. K., and Chen, K.-L. B. (2019). Using Big Data Analytics and Visualization to Create IoT-enabled Science Park Smart Governance Platform. In Nah, F. F.-H. and Siau, K., editors, *HCI in Business, Government and Organizations. Information Systems and Analytics*, pages 459–472, Cham. Springer International Publishing.

[274] Yang, R., Ford, B., Tambe, M., and Lemieux, A. (2014). Adaptive resource allocation for wildlife protection against illegal poachers. *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, 1(Aamas):453–460.

[275] Yazici, T. (2022). A proposal for the usage of reconnaissance satellites to monitor international human and wildlife trafficking hotspots. *Acta Astronautica*, 195:77–85.

[276] Ye, C. and Yang, Y. (2021). Wireless sensor network coverage algorithm based on deep forest. In *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture*, AIAM2021, page 1690–1694, New York, NY, USA. Association for Computing Machinery.

[277] Young, D. J., Koontz, M. J., and Weeks, J. (2022). Optimizing aerial imagery collection and processing parameters for drone-based individual tree mapping in structurally complex conifer forests. *Methods in Ecology and Evolution*, 13(7):1447–1463.

[278] Yuan, Y. (2021). *Reducing poaching risk through land use and patrol routes planning using data driven optimization*. PhD thesis, Carnegie Mellon University Pittsburgh, PA.

[279] Zafra-Calvo, N., Lobo, J., Prada, C., Nielsen, M., and Burgess, N. (2018). Predictors of elephant poaching in a wildlife crime hotspot: The ruvuma landscape of southern tanzania and northern mozambique. *Journal for Nature Conservation*, 41:79–87.

[280] Zehra, S., Mohsin, S. F. M., Wasi, S., Jami, S. I., Siddiqui, M. S., and Syed, M. K.-U.-R. R. (2021). Financial knowledge graph based financial report query system. *IEEE Access*, 9:69766–69782.

[281] Zheng, P., Xu, X., and Chen, C. H. (2020). A data-driven cyber-physical approach for personalised smart, connected product co-development in a cloud-based environment. *Journal of Intelligent Manufacturing*, 31(1):3–18.

[282] Zou, X. (2020). A survey on application of knowledge graph. *Journal of Physics: Conference Series*, 1487(1):012016.

[283] Zwerts, J. A., Stephenson, P., Maisels, F., Rowcliffe, M., Astaras, C., Jansen, P. A., van Der Waarde, J., Sterck, L. E., Verweij, P. A., Bruce, T., et al. (2021). Methods for wildlife monitoring in tropical forests: Comparing human observations, camera traps, and passive acoustic sensors. *Conservation Science and Practice*, 3(12):e568.

.1 Appendix I: Methodology Details

Developing and testing a linked data store -Forest Observatory (Activity Plan)
Introduction

Forest Observatory is a linked data store that integrates heterogeneous data from disparate sources and presents them in a unified manner. This project collaborates between the School of Computer Science and the School of Biosciences (and its Danau Girang Field Centre; DGFC) at Cardiff University. *Purpose of the activity* This activity is part of a PhD project, and the purpose is to collect information from bioscience researchers and environmental scientists to help develop the linked data store (Forest Observatory). **Proposed study structure:** We aim at conducting twelve to twenty interviews within two years. We will ask potential candidates from Danau Girange Centre (DGFC) and Biosciences (BIOSI) at Cardiff University to participate in our study. We will provide the nominated participant with the information sheet, demographic survey, and consent form. We will give them five working days to read the information sheet, ask questions about the study, and consider participation.

Once potential participants agreed and consent -unless they choose to withdraw at any point, we will schedule online 60-minute semi-structured recorded interviews for each one. We will use Zoom and Microsoft teams to conduct and record interviews plus an online whiteboarding application (e.g., Miro).

Proposed questions during the interview:

- What are the types of collected data?
- How do you process the collected data?
- What are the tools and methods used to process the data?
- How do you access and interact with the data?
- What are the drawbacks of your current data system?
- What are the questions that you require your data environment to answer?
- What would the ideal data model look like for you (e.g., chronological data catalogue,
- interactive interface with links to downloadable datasets)?

- What is your feedback about the delivered linked data store prototype/ outcome?

We require participants to provide demographic information listed on the attached survey, including three questions -with multiple choice answers - about their education level, occupation, and years range of experience. We confirm that the demographic survey provided is the complete list of questions.

Storing collected data All collected data will be stored securely through Cardiff University's OneDrive; we propose access to the data by: Academic staff: Professor Omer Rana and Dr Charith Perera. Research students: Naeima Hamed and Omar Mussa.

Demographic Survey

Thank you for filling out this form.

What is your level of education?

1. Some high school
2. High school graduate or equivalent
3. Trade or Vocational Degree
4. Some college
5. Associate degree
6. Bachelor's degree
7. Master's degree
8. Doctorate (PhD and DPhil)
9. Prefer not to answer

What is your occupation? How many years of experience do you have?

1. 1 to 3
2. 4 to 10
3. 11 to 20
4. more than 20

.2 Ontology Requirements Specification Document (ORSD)

.2.1 Purpose

The Forest Observatory Ontology (FOO) aims to describe wildlife digital data generated by sensors. The primary purpose is to backbone the Forest Observatory. That is, a linked datastore that allows unified access to heterogeneous wildlife data and enables standardised data exchange between different computer systems and applications.

.2.2 Scope

The Internet of Things (IoT) and wildlife make up the evolving scope of FOO. It adopts and combines classes and properties from Sensor Observation Sample and Actuation (SOSA) and BBC Wildlife Ontology (WO).

.2.3 Implementation Language

The Web Ontology Language (OWL2) is used to implement FOO.

.2.4 Intended End-Users

- Bioscientists.
- Wildlife Researchers.
- Computer Scientists.
- Data Scientists.

.2.5 Intended Uses

- To build linked data that offers data on-demand (i.e., granular data retrieval from disparate sources).
- For reasoning about the data of interest.
- Build Artificial Intelligence (AI) apps.

.2.6 Ontology Requirements

Functional Requirements

- FOO must include IoT elements, such as sensors.
- FOO must include wildlife concepts, such as taxon rank.

- FOO must contain the relationship between the Internet of Things (IoT) and wildlife concepts.
- 106 curated Competency Questions (CQs),
- 10 Natural Language Statements (NLSs), see table 1.

Non-Functional Requirements

- FOO must be scalable to accommodate increasing amounts of wildlife and IoT data.
- FOO should be interoperable with existing wildlife data standards and systems.
- FOO The ontology must perform efficiently when reasoning over large datasets.

.3 Competency Questions and their formulated SPARQL Queries

CQ 01: *Where does Elephant Jasmine forage?*

```
# Let the date be 2011-11-13
PREFIX geo:<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT * {
    ?observation  a  foo:gPSObservation ;
            foo:localDate    "2011-11-13"^^xsd:date ;
            geo:latitude   ?Latitude ;
            geo:longitude  ?longitude .}
```

Listing 1 SPARQL Query for Question 1

CQ 02: *What are the daily movement patterns for Elephant X in June?*

```
# Let elephant X be Abaw
PREFIX geo:<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT DISTINCT *
WHERE {
  ?observation a                foo:gPSObservation ;
              foo:madeBySensor  foo:abawGPS ;
              foo:localDate     ?date;
```

```
              geo:longitude     ?long;
              geo:latitude      ?lat.
FILTER(?date >= "2014-06-01"^^xsd:date && ?date <= "2014-06-30"^^xsd:date)
```

Listing 2 SPARQL Query for Question 2

CQ 03: *What are the yearly movement patterns for Elephant X?*

```
# Let elephant X be Puteri
PREFIX geo:<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT *{
  ?observation  a                foo:gPSObservation ;
              foo:madeBySensor   foo:puteriGPS ;
              foo:localDate      ?date ;
              geo:latitude       ?Latitude ;
              geo:longitude      ?longitude .
  FILTER(?date >= "2013-12-31"^^xsd:date && ?date <= "2014-12-31"^^xsd:date)}
```

Listing 3 SPARQL Query for Question 3

CQ 04: *How do the movements of Elephant X relate to human and urban areas?*

```
# Let human and urban areas be the oil Plam Plantation and elpehant X be Jasmin
PREFIX geo:<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT * {
  ?observation  a                     foo:gPSObservation ;
            foo:madeBySensor          foo:jasminGPS ;
            foo:hasFeatureOfInterest  foo:jasmin;
            foo:localDate             ?date ;
            geo:latitude              ?Latitude ;
            geo:longitude             ?longitude .
  FILTER(?date >= "2012-02-07"^^xsd:date && ?date <= "2012-02-15"^^xsd:date)}
```

Listing 4 SPARQL Query for Question 4

CQ 05: *Has elephant x died?*

```
# Let elephant x be Sandi
PREFIX geo:<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT DISTINCT * {
  ?observation  a                  foo:gPSObservation;
        foo:madeBySensor           foo:sandiGPS ;
        foo:hasFeatureOfInterest   foo:sandi ;
        foo:localDate              ?date;
        foo:cov                    ?cov;
        foo:speed                  ?speed.
   FILTER(?cov = "0.0"^^xsd:float && ?speed <= "0.1"^^xsd:float)}
```

Listing 5 SPARQL Query for Question 5

CQ 06: *Why did Elephant X die?*

```
# We do not know so we have to search near (10 Kilometer far) human domainated landscapes.
# This query works with the old modeling and using time ontology
SELECT DISTINCT * {
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX foo: <http://example.org/foo#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX unit: <http://qudt.org/vocab/unit#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
 SELECT * {?s            a        sosa:Observation;
                 time:gMTDate     "2011-11-13"^^xsd:date;
                 time:gMTTime     ?GMTTime;
                 pos:long         ?long;                          pos:lat          ?lat;
 geof:nearby (5.674 118.393  10 unit:Kilometer). #Assuming the oil palm pantation
 is located with 10 Kilometer from this coordinates.
        ?s1
                pos:long        ?long1;
                pos:lat         ?lat1;
                time:gMTDate    "2011-11-13"^^xsd:date;
 BIND (geof:distance(?s, ?s1, unit:Kilometer) as ?Distance) }
```

Listing 6 SPARQL Query for Question 6

CQ 07: *What are the optimal environmental conditions for Elephant X to survive?*

```
 # Elephants require large amounts of land to thrive and meet their ecological demands,
 which include food, water, and space.  Using GPS collars, behaviour and site sampling
 help to understand recursion to forging sites. Elephants revisit certain sites for their
 greater value.  The amount of time animals spend at these locations and how frequently
 they return to them can help us understand habitat quality and its value to animals, or
 individual resource quality and its importance within foraging sites.  If an animal
 returns to a spot and spends more time there than at other sites, this may help to
 locate high-quality areas or more vital resources. Longer time spent at a site.
 However, recursion behaviour may be a signal of deteriorating acceptable habitat
 quality or capacity since, as prime habitat becomes less available, recursion frequency
 should increase while time spent at locations decreases. * elephants preferred food plants
 like grass and bamboos.  Foraging area focus as per English et al. on the areas between
 Abai and Batu Puteh (5.18-N 5 42'N, 117.54-E 118.33-E), which were the downriver and upriver
 limits of the LKWS elephant population's range. The study area also contains 7 sections,
 each section refereed to as a 'lot' (approximately 218 km), including 89KM of protected
 forest reserves. The elephant herds utilised their whole range throughout the year including
 the use of privately owned forests and cultivated land, particularly oil palm plantations
 that were adjacent to and between forested areas.
 Another answer is protected areas in Southeast Sabah.

PREFIX foo: <http://w3id.org/def/foo#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX unit: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>


SELECT ?s (geof:distance(?geom, ?targetGeom, unit:Meter) AS ?Distance)
 {
  ?s       a         foo:gPSObservation;
           foo:long  ?long;
           foo:lat   ?lat.
  BIND(STRDT(CONCAT('POINT(', STR(?long), ' ', STR(?lat), ')'), geo:wktLiteral) AS ?geom)
  BIND(STRDT('POINT(117.54 5.18)', geo:wktLiteral) AS ?targetGeom)
  FILTER(geof:distance(?geom, ?targetGeom, unit:Meter) < 10000)
}
```

Listing 7 SPARQL Query for Question 7

CQ 08: *What can we learn from the movements of Elephants X, Y, and Z?*

```
#  Let Elephants X, Y, and Z be Aqeela, Ita, and Dara. Querying their
geo-locations in a unified manner allows us to learn about their movements.
PREFIX foo:<https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT * {
  {?AqeelaGPS  a  foo:gPSObservation;
                pos:longitude ?longX;
                pos:latitude ?latX;
                 ?predicateX ?elephantAqeela.
    FILTER(?elephantAqeela = <https://w3id.org/def/foo#aqeela>)
   }
#   UNION
  {
   ?ItaGPS  a  foo:gPSObservation;
                pos:longitude ?longY;
                pos:latitude ?latY;
                 ?predicateY ?elephantIta.
    FILTER(?elephantIta = <https://w3id.org/def/foo#ita>)
   }
  UNION
  {
   ?DaraGPS  a  foo:gPSObservation;
                pos:longitude ?longZ;
                pos:latitude ?latZ;
                 ?predicateZ ?elephantDara.
    FILTER(?elephantDara = <https://w3id.org/def/foo#dara>)
   }}
```

Listing 8 SPARQL Query for Question 8

CQ 09: *How does Elephant X use Habitat Site Y?*

```
#According to English et al., 2014, the site is defined as the area covering a
100m radius surrounding each measurement point taken from the center of
the elephant herd.
#Let elephant X be Aqeela
PREFIX foo:<https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX unit: <http://www.opengis.net/def/uom/OGC/1.0/>
SELECT *
{
```

```
  ?observation  a                          foo:gPSObservation;
              foo:madeBySensor             ?elephantGPS ;
              foo:hasFeatureOfInterest     ?elephant ;
              pos:longitude                ?Longitude ;
              pos:latitude                 ?Latitude.
BIND(STRDT(CONCAT("POINT(", STR(?Longitude), "␣", STR(?Latitude), ")"),
geo:wktLiteral) AS ?observationPoint)
BIND(STRDT("POINT(118.3019␣5.510)", geo:wktLiteral) AS ?referencePoint)
FILTER(geof:distance(?observationPoint, ?referencePoint, unit:metre) < 100)}}
```

Listing 9 SPARQL Query for Question 09

CQ 10: *What is the range of habitat sites used by Elephants X, Y, and Z?*

```
# let Elephants X Y, Z be Ita, Abaw and Jasmin
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT *
{
  ?observation a                    foo:gPSObservation ;
               foo:madeBySensor     ?gpsCollar   ;
               geo:latitude         ?latitude ;
               geo:longitude        ?longitude .
FILTER ( ?gpsCollar IN (foo:jasminGPS, foo:abawGPS, foo:itaGPS ))}
```

Listing 10 SPARQL Query for Question 10

CQ 11: *Where was Elephant X located during the flood season in the Lower Kinabatangan area?*

```
# Note: Flooding occurs between November and March during the west monsoon.
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT DISTINCT *
{
  ?observation a                        foo:gPSObservation;
              foo:madeBySensor          ?elephantGPS ;
              foo:hasFeatureOfInterest  ?elpehant ;
              foo:localDate             ?date ;
              pos:longitude             ?long ;
              pos:latitude              ?lat  .
  FILTER(?date >= "2011-11-01"^^xsd:date && ?date <= "2012-03-30"^^xsd:date)}
```

Listing 11 SPARQL Query for Question 11

CQ 12: *What was the average speed of Elephant X during the flood season?*

```
# Note: Flooding occurs between November and March during the west monsoon.
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT (AVG(?Speed) AS ?AVGspeed)
{
  ?observation a        foo:gPSObservation;
             foo:speed  ?speed ;
             foo:date   ?date .
  FILTER (?date >= "2012-02-07"^^xsd:date && ?date < "2012-02-15"^^xsd:date)
}
```

Listing 12 SPARQL Query for Question 12

CQ 13: *Is Elephant Dara near (5 Km) the danger zone (poachers' area) today?*

```
# Let poacher area be POINT(117.30193 5.510).
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX unit: <http://qudt.org/vocab/unit#>
SELECT DISTINCT *
{
?observation a             foo:gPSObservation;
             pos:longitude ?Longitude;
             pos:latitude  ?Latitude.
# Define observation and reference points as WKT literals
BIND(STRDT(CONCAT("POINT(", STR(?Longitude), "␣", STR(?Latitude), ")"),
geo:wktLiteral) AS ?observationPoint)
BIND("POINT(117.30193␣5.510)"^^geo:wktLiteral AS ?referencePoint)
# Apply distance filter to select observations within 10 km of the reference point
FILTER(geof:distance(?observationPoint, ?referencePoint, unit:Kilometer) < 5)}
```

Listing 13 SPARQL Query for Question 13

CQ 14: *How did Elephant X's movements change with climate change in 2014?*

```
# We need weather data to answer this question.
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
SELECT *
{
  # Filter for Elephant X's movement observations in 2014
  ?observation a                    foo:gPSObservation ;
               foo:localDate        ?date ;
               pos:latitude         ?latitude ;
               pos:longitude        ?longitude .
FILTER (?date >= "2014-01-01"^^xsd:date && ?date <= "2014-12-31"^^xsd:date)
#   # Link climate data with spatial and temporal proximity to Elephant X's observations
#   ?climateObservation a           climate:ClimateObservation ;
#              climate:temperature   ?temperature ;
#              climate:precipitation ?precipitation ;
#              foo:localTime         ?date ;
#              pos:lat               ?latitude ;
#              pos:long              ?longitude .
}
# ORDER BY ?date
```

Listing 14 SPARQL Query for Question 14

CQ 15: *What are Elephant X's preferred habitats based on prolonged stays in areas?*

```
# Foraging area focus as per English et al. on the areas
#between Abai and Batu Puteh (5.18-N 5 42'N, 117.54-E 118.33-E),
# which were the downriver and upriver limits of the LKWS elephant population's range.
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?location ?latitude ?longitude (COUNT(?date) AS ?stayDuration)
{
  ?observation a                    foo:gPSObservation ;
               foo:FeatureOfInterest ?elephant ;
               foo:date              ?date ;
               pos:latitude          ?latitude ;
               pos:longitude         ?longitude .
  # Filter for locations within the specified latitude and longitude range
  FILTER (?latitude >= "5.18"^^xsd:float && ?latitude <= "5.42"^^xsd:float)
  FILTER (?longitude >= "117.54"^^xsd:float && ?longitude <= "118.33"^^xsd:float)
  # Generate a location identifier (combination of lat and long) for grouping
  BIND(CONCAT(STR(?latitude), ",", STR(?longitude)) AS ?location)
```

```
 }
GROUP BY ?location ?latitude ?longitude
HAVING (?stayDuration >= 0)
ORDER BY DESC(?stayDuration)
```

Listing 15 SPARQL Query for Question 15

CQ 16: *How far was Elephant X from the oil plantation fencing?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX unit: <http://qudt.org/vocab/unit#>
SELECT DISTINCT ?s ?plantationLocation (geof:distance(?geo1,
?plantationLocation, unit:Kilometer) AS ?Distance)
 {
  ?s      a                    foo:gPSObservation;
          pos:longitude        ?long;
          pos:latitude         ?lat;
          foo:localDate        ?date.
# Bind geographic points for each observation
 BIND(STRDT(CONCAT("POINT(", STR(?long), "␣", STR(?lat), ")"), geo:wktLiteral) AS ?geo1)
# Define the fixed location point for the plantation
 BIND("POINT(118.393␣5.674)"^^geo:wktLiteral AS ?plantationLocation)
# Filter distances to be within 5 kilometers of the plantation location
FILTER (geof:distance(?plantationLocation, ?geo1, unit:Kilometer) <= 5)}
```

Listing 16 SPARQL Query for Question 16

CQ 17: *When was Elephant X near the oil plantation fencing?*

```
 # Let elephant X be Ita.
PREFIX foo:<https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX unit: <http://qudt.org/vocab/unit#>
SELECT *
{
  ?s  a                        foo:gPSObservation ;
```

```
     foo:madeBySensor           foo:itaGPS;
     foo:hasFeatureOfInterest   foo:ita;
     foo:localDate              ?localDate;
     foo:localTime              ?localTime;
     pos:longitude              ?long;
     pos:latitude               ?lat.
# Construct the observation point as WKT literal
BIND(STRDT(CONCAT("POINT(", STR(?long), " ", STR(?lat), ")"), geo:wktLiteral)
AS ?observationPoint)
# Define the fixed reference point
BIND("POINT(118.393 5.674)"^^geo:wktLiteral
AS ?referencePoint)
# Calculate the distance
BIND(geof:distance(?observationPoint, ?referencePoint, unit:Kilometer) AS ?Distance)
# Filter to within 5 km distance
FILTER(?Distance <= 5)}
```

Listing 17 SPARQL Query for Question 17

CQ 18: *What is the distance traveled between each of Elephant X's stops (sleeping)? (Query took 184930 ms*

```
# To calculate the distance traveled between each of Elephant X's stops,
# filtering for stops where "sleeping" or prolonged pauses are identified
# based on a significant time difference between consecutive observations.
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX unit: <http://qudt.org/vocab/unit#>
PREFIX foo: <https://w3id.org/def/foo#>
SELECT ?prevStop ?prevDate ?nextStop ?nextDate ?distanceTraveled{
  # Get observations for Elephant ita with timestamp, lat, long
  ?prevStop a foo:gPSObservation ;
            foo:hasFeatureOfInterest foo:ita ;
            foo:localDate ?prevDate ;
            pos:longitude ?prevLong ;
            pos:latitude ?prevLat .
  ?nextStop a foo:gPSObservation;
            foo:hasFeatureOfInterest foo:ita ;
            foo:localDate ?nextDate ;
            pos:longitude ?nextLong ;
            pos:latitude ?nextLat .
```

```
␣␣#␣Observations␣are␣ordered␣and␣calculate␣only␣for␣consecutive␣stops␣with␣a␣longer␣time␣gap
␣␣FILTER␣(?nextDate␣>␣?prevDate)
␣␣#␣Calculate␣the␣distance␣between␣consecutive␣stops
␣␣BIND(STRDT(CONCAT("POINT(",␣STR(?prevLong),␣"␣",␣STR(?prevLat),␣")"),␣geo:wktLiteral)
␣␣AS␣?prevLocation)
␣␣BIND(STRDT(CONCAT("POINT(",␣STR(?nextLong),␣"␣",␣STR(?nextLat),␣")"),␣geo:wktLiteral)
␣␣AS␣?nextLocation)
␣␣BIND(geof:distance(?prevLocation,␣?nextLocation,␣unit:Kilometer)␣AS␣?distanceTraveled)}
ORDER␣BY␣?prevDate
```

Listing 18 SPARQL Query for Question 18

CQ 19: *Which elephants met this month?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo:<https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?ElephantX ?ElephantY ?dateX ?dateY
WHERE {
  ?ObservationX a foo:gPSObservation;
                foo:hasFeatureOfInterest   ?ElephantX ;
                foo:localDate ?dateX;
                pos:longitude ?longX;
                pos:latitude ?latX.

  ?ObservationY a  foo:gPSObservation;
                foo:hasFeatureOfInterest   ?ElephantY ;
                foo:localDate ?dateY;
                pos:longitude ?longY;
                pos:latitude ?latY.

  FILTER(?ElephantX != ?ElephantY)
  FILTER(MONTH(xsd:date(?dateX)) = MONTH(NOW()) && MONTH(xsd:date(?dateY)) = MONTH(NOW()))
  # FILTER(?dateX = ?dateY)
  FILTER(?longX = ?longY && ?latX = ?latY)}
```

Listing 19 SPARQL Query for Question 19

CQ 20: *Which sites were revisited by Elephant X month?*

```
# Let elephant X be Abaw
PREFIX foo: <https://w3id.org/def/foo#>
```

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?location (COUNT(?location) AS ?visits)
# SELECT *
WHERE {
  ?observation a  foo:gPSObservation;
              foo:hasFeatureOfInterest  foo:abaw;
              foo:localDate ?date;
              pos:longitude ?long;
              pos:latitude ?lat.

  # Constructing a simple identifier for a ""location"" based on its long/lat.
BIND(CONCAT(STR(?long), "-", STR(?lat)) AS ?location)
FILTER(MONTH(xsd:date(?date)) = MONTH(NOW()) && YEAR(xsd:date(?date)))}
GROUP BY ?location
HAVING (COUNT(?location) > 1)
```

Listing 20 SPARQL Query for Question 20

CQ 21: *What environment or habitat does Elephant X prefer, based on the prolonged time spent in a certain area?*

```
# SPARQL query can be crafted that identifies the specific types of environments
# where Elephant X spends extended periods- focusing
# on areas where the time spent between consecutive observations exceeds a threshold.
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX unit: <http://qudt.org/vocab/unit#>
PREFIX foo: <https://w3id.org/def/foo#>
SELECT ?location ?prevDate ?nextDate ?timeSpent
WHERE {
  # Get GPS observations for Elephant Ita with timestamp, lat, long
  ?obs1 a foo:gPSObservation ;
        foo:hasFeatureOfInterest foo:ita ;
        foo:localDate ?prevDate ;
        pos:longitude ?prevLong ;
        pos:latitude ?prevLat .

  ?obs2 a foo:gPSObservation ;
        foo:hasFeatureOfInterest foo:ita ;
        foo:localDate ?nextDate ;
```

```
        pos:longitude ?nextLong ;
        pos:latitude ?nextLat .
  # Ensure observations are ordered and calculate time spent in the area
  FILTER (?nextDate > ?prevDate)
  BIND((?nextDate - ?prevDate) AS ?timeSpent)
  # Identify prolonged stays
  FILTER(?timeSpent >= "PT8H"^^xsd:duration)  # Adjust duration threshold as necessary
  # Calculate location
  BIND(STRDT(CONCAT("POINT(", STR(?prevLong), "␣", STR(?prevLat), ")"), geo:wktLiteral)
  AS ?location) }
ORDER BY DESC(?timeSpent)
```

Listing 21 SPARQL Query for Question 21

CQ 22: *Was there any significant change in Elephant X's movement patterns between June and July 2012?*

```
 # Elephant X was Putut
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?ElephantX ?year ?month
       (AVG(?speed) AS ?avgSpeed)
       (SUM(?distance) AS ?totalDistance)
       (AVG(?direction) AS ?avgDirection)
WHERE {
   ?ObservationX a foo:gPSObservation;
                 foo:hasFeatureOfInterest ?ElephantX;
                 foo:localDate ?date;
                 pos:latitude ?latitude;
                 pos:longitude ?longitude;
                 foo:speed ?speed;
                 foo:direction ?direction;
                 foo:distance ?distance .
   BIND(YEAR(?date) AS ?year)
   BIND(MONTH(?date) AS ?month)

   FILTER ((?date >= "2012-06-01"^^xsd:date && ?date <= "2012-07-31"^^xsd:date))}
GROUP BY ?ElephantX ?year ?month
ORDER BY ?year ?month
```

Listing 22 SPARQL Query for Question 22

CQ 23: *Has Elephant X visited Village Y in year Z?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?ElephantX (COUNT(?ObservationX) AS ?visits)
WHERE {
  ?ObservationX a foo:gPSObservation;
                foo:hasFeatureOfInterest ?ElephantX;
                foo:localDate ?dateX;
                pos:longitude ?longX;
                pos:latitude ?latX.

  # Geographic bounds for Village Y
  FILTER (?longX >= 118.0000 && ?longX <= 118.8333 && ?latX >= 5.3333 && ?latX <= 5.8333)

  # Filter for observations in the year 2012
  FILTER (YEAR(?dateX) = 2012)}
GROUP BY ?ElephantX
```

Listing 23 SPARQL Query for Question 23

CQ 24: *What is the movement range of Elephant X during Month Y?*

```
# Let Elephant X be Jasmin and Month X be June 2011
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
{
  ?ObservationX a foo:gPSObservation;
                foo:hasFeatureOfInterest foo:jasmin;
                foo:localDate ?dateX;
                pos:longitude ?longX;
                pos:latitude ?latX.
  FILTER ((?dateX >= "2011-11-01"^^xsd:date) && (?dateX <= "2012-11-30"^^xsd:date))}
```

Listing 24 SPARQL Query for Question 24

CQ 25: *What is Elephant's activity (speed) during Month Y?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?ElephantX ?dateX (AVG(?speed) AS ?averageSpeed)
{
  ?ObservationX a foo:gPSObservation;
                foo:hasFeatureOfInterest ?ElephantX;
                foo:localDate ?dateX;
                foo:speed ?speed .
  FILTER ((?dateX >= "2011-11-01"^^xsd:date) && (?dateX <= "2011-11-30"^^xsd:date))}
GROUP BY ?ElephantX ?dateX
```

Listing 25 SPARQL Query for Question 25

CQ 26: *Are there any interactions between collared elephants during the flood season?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephantA ?elephantB ?longA ?latA ?longB ?latB ?DateA ?DateB
WHERE {
  ?obsA a                         foo:gPSObservation ;
        foo:localDate             ?DateA;
        foo:hasFeatureOfInterest  ?elephantA;
        pos:longitude             ?longA;
        pos:latitude              ?latA.
  ?obsB  a                        foo:gPSObservation ;
        foo:localDate             ?DateB;
        foo:hasFeatureOfInterest  ?elephantB;
        pos:longitude             ?longB;
        pos:latitude              ?latB.
  # Ensure we're␣considering␣two␣different␣elephants
␣␣FILTER␣(?elephantA␣!=␣?elephantB)
␣␣#␣#␣Date␣range␣filter␣for␣observations␣within␣flood␣season␣(NOV␣2011␣-␣MARCH␣2012)
␣␣#␣FILTER␣(?DateA␣>=␣"2013-11-01"^^xsd:dateTime␣&&␣?DateB␣>=␣"2013-11-01"^^xsd:dateTime)
␣␣#␣#␣Filter␣for␣interactions␣within␣a␣day
␣␣#␣FILTER␣(ABS(?DateA␣-␣?DateB)␣<=␣"P1D"^^xsd:duration)
␣␣#␣Proximity␣filter:␣Approximate␣spatial␣interaction␣(adjusting␣to␣~5km␣radius)
```

```
  FILTER (ABS(?longA - ?longB) <= 0.05 && ABS(?latA - ?latB) <= 0.05)
}
# ORDER BY ?DateA ?DateB
```

Listing 26 SPARQL Query for Question 26

CQ 27: *What is the status of Elephant X's tracking collar battery?*

```
# Let elephant X be Ita
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
 {
  ?observation foo:hasFeatureOfInterest foo:ita;  # Linking to Elephant Ita
               foo:madeBySensor        ?GPScollar;
               foo:localDate           ?date ;
               foo:speed               ?speed .
  # Filter for observations where speed is zero
  FILTER(?speed = 0)}
```

Listing 27 SPARQL Query for Question 27

CQ 28: *What habitat has Elephant X selected this season?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?habitat (COUNT(?observation) AS ?visits)
WHERE {
  ?observation a foo:gPSObservation;
                foo:hasFeatureOfInterest  ?elpehantx;
                foo:localDate ?date.

  ?location a   foo:treeObservation ;
                foo:site ?habitat ;
                foo:date  ?sitedate. }
GROUP BY ?habitat
ORDER BY DESC(?visits)
LIMIT 1
```

Listing 28 SPARQL Query for Question 28

CQ 29: *What is the average elevation of Elephant X during a specific time range?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
# (AVG(?altitude) AS ?averageElevation)
WHERE {
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest  ?elephantX;
               foo:localDate ?date;
               foo:altitude ?altitude.

# Filter for observations within the specified time range (November 2011 to March 2012)
  FILTER (?date >= "2011-11-01"^^xsd:date && ?date <= "2012-03-31"^^xsd:date)}
LIMIT 1
```

Listing 29 SPARQL Query for Question 29

CQ 30: *Which elephant came near the logged site?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?elephantX
WHERE {
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest  ?elephantX;
               pos:latitude ?lat;
               pos:longitude ?long.
  FILTER (ABS(?lat - 5.0) < 0.01 && ABS(?long - 118.0) < 0.01)
  ## Assume it is the logged site location.}
```

Listing 30 SPARQL Query for Question 30

CQ 31: *Which elephant came near the semi-logged site?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
```

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT *
WHERE {
  # GPS sensor for elephant
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephantX;
               pos:latitude ?lat;
               pos:longitude ?long.

  # Soil Sensor for land use information (we need geo-location)
  ?soilObservation a foo:soilObservation;
                   foo:landUse ?landUse.
  # Filter for land use exactly 'logged'
  FILTER (?landUse = "semi-logged")}
```

Listing 31 SPARQL Query for Question 31

CQ 32: *Which elephants crossed the river?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?elephant
{
  ?observationWest a foo:gPSObservation;
                   foo:madeBySensor  ?elephant;
                   pos:latitude ?latWest;
                   pos:longitude  ?longWest.
  ?observationEast a foo:gPSObservation;
                   foo:madeBySensor ?elephant;
                   pos:latitude ?latEast;
                   pos:longitude ?longEast.
  FILTER(?longWest < 118.1 && ?longEast > 118.1)}
```

Listing 32 SPARQL Query for Question 32

CQ 33: *What is the canopy height for the distance traveled by Elephant X during the flood season?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?date ?location (AVG(?canopyHeight) AS ?averageCanopyHeight)
```

```
{
  ?observation a foo:gPSObservation;
               foo:madeBySensor ?elephant;
               foo:localDate ?date;
               pos:latitude ?lat;
               pos:longitude ?long.
  ?vegetation foo:treeHeight_m ?canopyHeight;
  # Adjusted filter for the flood season spanning from November to March
  FILTER ((?date >= "2011-11-01"^^xsd:date || ?date <= "2012-03-31"^^xsd:date))
  # Bind the latitude and longitude as a single string for location
  BIND(CONCAT(STR(?lat), ", ", STR(?long)) AS ?location)}
GROUP BY ?date ?location
ORDER BY ?date
```

Listing 33 SPARQL Query for Question 33

CQ 34: *Which elephants are near the oil palm plantations this week?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?elephant {
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude ?elephantLat;
               pos:longitude ?elephantLong .
   foo:OilPalmPlantation a  owl:Class; # we need reasoning or plantation location.
                #  pos:latitude ?plantationLat;
                #  pos:longitude ?plantationLong.
  # # Proximity filter within ~11 km (0.1 degrees)
  # FILTER (ABS(?elephantLat - ?plantationLat) < 0.1 &&
  ABS(?elephantLong - ?plantationLong) < 0.1)
  # # Date filter for the specific week (October 2 to October 8, 2023)
  # FILTER (?observationTime >= "2011-10-02"^^xsd:date
  && ?observationTime <= "2011-10-08"^^xsd:date)}
```

Listing 34 SPARQL Query for Question 34

CQ 35: *What is the home range for all collared elephants?*

```
# 35. What is the home range for all collared elephants?
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephant (SAMPLE(?lat) AS ?latitude) (SAMPLE(?long) AS ?longitude)
(COUNT(?observation) AS ?observations)
WHERE {
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               foo:madeBySensor ?GPS;
               pos:latitude ?lat;
               pos:longitude ?long.}
GROUP BY ?elephant
```

Listing 35 SPARQL Query for Question 35

CQ 36: *What is the distance traveled by Elephant Y over a specific period?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?elephant ?date {
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               foo:localDate ?date;
               pos:latitude ?elephantLat;
               pos:longitude ?elephantLong .
  FILTER (?date >= "2012-01-01"^^xsd:date && ?date <= "2012-01-31"^^xsd:date)}
ORDER BY ?date
```

Listing 36 SPARQL Query for Question 36

CQ 37: *What are the altitudes of the collared elephants?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephant (MAX(?date) AS ?latestObservationDate)
(SAMPLE(?altitude) AS ?latestAltitude)
{
  ?observation a                        foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
```

```
              foo:madeBySensor         ?GPS;
              foo:localDate            ?date ;
              foo:altitude             ?altitude. }
 GROUP BY ?elephant
```

Listing 37 SPARQL Query for Question 37

CQ 38: *What are the body/environment temperatures for collared elephants?*

```
# 38. What are the body/environment temperatures for collared elephants?
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephant (MAX(?date) AS ?latestObservationDate) (SAMPLE(?bodyTemp)
AS ?latestBodyTemperature) WHERE {
  ?observation a  foo:gPSObservation;
                  foo:madeBySensor  ?sensor;
                  foo:localDate  ?date ;
                  foo:gMTDate    ?gMTDate  ;
                  foo:temperature  ?bodyTemp ;
                  foo:hasFeatureOfInterest  ?elephant.
}
GROUP BY ?elephant
ORDER BY ?elephant
```

Listing 38 SPARQL Query for Question 38

CQ 39: *What is the behavior of Elephants X and Y this month?*

```
# Let Elephant X be Bikang1 and Elephant Y be Bikang2
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?observationX a                      foo:gPSObservation;
              foo:speed                ?speedX;
              foo:distance             ?distanceX;
              pos:longitude            ?longX;
              pos:latitude             ?latX;
              foo:localDate            ?observationTimeX ;
              foo:hasFeatureOfInterest foo:bikang1.
  ?observationY a                      foo:gPSObservation;
```

```
              foo:speed                ?speedY;
              foo:distance             ?distanceY;
              pos:longitude            ?longY;
              pos:latitude             ?latY;
              foo:localDate            ?observationTimeY ;
              foo:hasFeatureOfInterest foo:bikang2.
 FILTER (?observationTimeX >= "2013-05-01"^^xsd:date &&
          ?observationTimeX <= "2013-05-30"^^xsd:date
          && ?observationTimeY>= "2013-05-01"^^xsd:date &&
          ?observationTimeY <= "2013-05-30"^^xsd:date)}
```

Listing 39 SPARQL Query for Question 39

CQ 40: *Does Elephant X need help?*

```
# Let Elephant X be Ita
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT (SAMPLE(?speed) AS ?latestValue) WHERE {
  ?observation a                        foo:gPSObservation;
              foo:hasFeatureOfInterest  foo:ita ;
              foo:localTime             ?time ;
              foo:speed                 ?speed.}
```

Listing 40 SPARQL Query for Question 40 label

CQ 41: *What are the distribution patterns of Elephants X and Y during this month?*

```
 # Let Elephant X be Dara and Elephant Y be Kuma
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?observationX a                     foo:gPSObservation;
              pos:longitude           ?longX;
              pos:latitude            ?latX;
              foo:localDate           ?observationTimeX ;
              foo:hasFeatureOfInterest foo:dara.
  ?observationY a foo:gPSObservation;
              pos:longitude            ?longY;
              pos:latitude             ?latY;
```

```
              foo:localDate            ?observationTimeY ;
              foo:hasFeatureOfInterest foo:kuma.
 FILTER (?observationTimeX >= "2013-05-01"^^xsd:date &&
          ?observationTimeX <= "2013-05-30"^^xsd:date &&
          ?observationTimeY>= "2013-05-01"^^xsd:date &&
          ?observationTimeY <= "2013-05-30"^^xsd:date)}
```

Listing 41 SPARQL Query for Question 41 label

CQ 42: *Are Elephants X and Y's favorite foods in a particular area?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?observationX a                      foo:gPSObservation;
              pos:longitude            ?longX;
              pos:latitude             ?latX;
              foo:localDate            ?observationTimeX ;
              foo:hasFeatureOfInterest ?ElephantX.
  ?observationY a                       foo:gPSObservation;
              pos:longitude             ?longY;
              pos:latitude              ?latY;
              foo:localDate             ?observationTimeY ;
              foo:hasFeatureOfInterest  ?ElephantY.
#  Bornean elephants look for food near oil palm plantations in Sabah, Malaysia,
FILTER (?lat >= 4.23 && ?lat <= 5.32 && ?long >= 117.23 && ?long <= 118.40)}
```

Listing 42 SPARQL Query for Question 42 label

CQ 43: *Do we need to create corridors along rivers/palm plantations, or is it not an obstacle for elephants to cross the river?*

```
# [NO SPARQL QUERY]: The decision to create corridors should be based
on a multidisciplinary approach, incorporating data-driven insights
from SPARQL queries and GIS analysis with on-the-ground ecological
knowledge and conservation strategies.
```

Listing 43 SPARQL Query for Question 43 label

CQ 44: *Why have the elephants' collars been fitted for almost two years?*

```
# P2062D is approximately 5 years and 238 days.
#Therefore, the duration of fitted collars are more than two years.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT
  (MIN(?collarDate) AS ?oldestDate)
  (MAX(?collarDate) AS ?latestDate)
  ((MAX(?collarDate) - MIN(?collarDate)) AS ?duration)
{
  ?observation a foo:gPSObservation;
               foo:localDate ?collarDate.}
```

Listing 44 SPARQL Query for Question 44 label

CQ 45: *What are the migration patterns of Elephants X during the flood season?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest  ?elephant ;
               pos:longitude ?long;
               pos:latitude ?lat;
               foo:localDate ?date ;
  # Filter observations to the flood season period
  # flood season spans from November to March
 FILTER (
    (?date >= "2011-11-01"^^xsd:date && ?date <= "2011-12-31"^^xsd:date) ||
    (?date >= "2012-01-01"^^xsd:date && ?date <= "2012-03-31"^^xsd:date))
}
ORDER BY ?date
```

Listing 45 SPARQL Query for Question 45 label

CQ 46: *What are the favorite locations that Elephant X likes to visit during certain times of the year?*

```
# Let Elephant X be Jasmin
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?location (COUNT(?observation) AS ?visits)
 {?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest  foo:jasmin ;
               pos:longitude ?long;
               pos:latitude ?lat;
               foo:localDate ?date ;
# Construct a simple identifier for a ""location"" based on lat/long
  BIND(CONCAT(STR(?lat), "","", STR(?long)) AS ?location)
# Filter for a specific time frame, e.g., summer months
  FILTER (?date >= "2012-06-01"^^xsd:date && ?date <= "2012-06-30"^^xsd:date)}
GROUP BY ?location
ORDER BY DESC(?visits)
```

Listing 46 SPARQL Query for Question 46 label

CQ47: *Where are elephants likely to come into contact with humans?*

```
# According to research, it's␣near␣oil␣palm␣planations.
PREFIX␣foo:␣<https://w3id.org/def/foo#>
PREFIX␣pos:␣<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX␣xsd:␣<http://www.w3.org/2001/XMLSchema#>
PREFIX␣owl:␣<http://www.w3.org/2002/07/owl#>
SELECT␣?elephant␣?elephantLat␣?elephantLong␣?plantationLat␣?plantationLong␣?comment{
␣␣#␣Elephant␣locations
␣␣?observation␣a␣foo:gPSObservation;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣foo:hasFeatureOfInterest␣␣?elephant;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣pos:latitude␣?elephantLat;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣pos:longitude␣?elephantLong.
␣␣#␣Oil␣Palm␣Plantation␣details
␣foo:OilPalmPlantation␣a␣owl:Class;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣rdfs:comment␣?comment.
␣␣#␣Assign␣hypothetical␣coordinates␣for␣the␣plantation
␣␣(since␣exact␣locations␣aren't shared)
  BIND(5.6 AS ?plantationLat)
  BIND(118.1 AS ?plantationLong)
  # Filter to match within the same geographic area
  FILTER (?elephantLat >= 5.24 && ?elephantLat <= 5.76 &&
  ?elephantLong >= 117.54 && ?elephantLong <= 118.86)
  FILTER (?plantationLat >= 5.24 && ?plantationLat <= 5.76 &&
```

```
  ?plantationLong >= 117.54 && ?plantationLong <= 118.86)}
```

Listing 47 SPARQL Query for Question 47 label

CQ48: *What are the places where elephants may be vulnerable?*

```
# Query for areas of human-elephant conflict (within 5Km radius from oil palm plantation)
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT ?elephant ?elephantLat ?elephantLong ?plantationLat ?plantationLong ?comment
{
  # Elephant locations
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude ?elephantLat;
               pos:longitude ?elephantLong.
  # Oil Palm Plantation details
  foo:OilPalmPlantation a owl:Class;
                        rdfs:comment ?comment.
  # Assign hypothetical coordinates for the plantation (since exact locations aren't␣shared)
␣␣BIND(5.36␣AS␣?plantationLat)
␣␣BIND(118.66␣AS␣?plantationLong)
␣␣#␣Approximate␣distance␣calculation␣(flat-earth␣approximation)
␣␣BIND(
␣␣␣␣SQRT(
␣␣␣␣␣␣POW((?elephantLat␣-␣?plantationLat)␣*␣111.32,␣2)
␣␣␣␣␣␣+␣POW((?elephantLong␣-␣?plantationLong)␣*␣111.32,␣2)
␣␣␣␣)␣AS␣?distance
␣␣)
␣␣FILTER␣(?distance␣<=␣5)␣#␣Keep␣only␣results␣within␣5␣km␣radius
␣}
```

Listing 48 SPARQL Query for Question 48 label

CQ49: *Where can we assign locations to rangers?*

```
# It's␣recommended␣to␣deploy␣rangers␣whree␣there␣are␣elephants␣herds.␣␣␣
PREFIX␣foo:␣<https://w3id.org/def/foo#>
PREFIX␣pos:␣<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX␣xsd:␣<http://www.w3.org/2001/XMLSchema#>
```

```
SELECT␣?rangerLocation␣?centerLat␣?centerLong␣(COUNT(?elephant)␣AS␣?elephantCount)
WHERE␣{
␣␣#␣Elephant␣locations
␣␣?observation␣a␣foo:gPSObservation;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣foo:hasFeatureOfInterest␣?elephant;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣pos:latitude␣?lat;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣pos:longitude␣?long.

␣␣#␣Group␣elephant␣observations␣by␣clusters␣(assuming␣predefined␣ranger␣zones)
␣␣BIND(FLOOR(?lat␣*␣10)␣/␣10␣AS␣?centerLat)␣#␣Approximate␣grouping␣by␣latitude
␣␣BIND(FLOOR(?long␣*␣10)␣/␣10␣AS␣?centerLong)␣#␣Approximate␣grouping␣by␣longitude
␣␣BIND(CONCAT(STR(?centerLat),␣",",␣STR(?centerLong))␣AS␣?rangerLocation)
}
GROUP␣BY␣?rangerLocation␣?centerLat␣?centerLong
ORDER␣BY␣DESC(?elephantCount)
LIMIT␣10
```

Listing 49 SPARQL Query for Question 49 label

> **CQ50:** *How to track (investigate) the last location of a dead elephant?*

```
# if the speed is zero for long time
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephant ?lat ?long ?locationCluster (COUNT(?observation) AS ?durationCount){
  # Observations of elephant locations
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude ?lat;
               pos:longitude ?long;
               foo:localDate ?observationTime;
               foo:speed ?speed.
  # Filter for zero speed
  FILTER(?speed = 0)
  # Group observations by the same location to track "no␣movement"
  BIND(CONCAT(STR(?lat), ",", STR(?long)) AS ?locationCluster)
  }
GROUP BY ?elephant ?lat ?long ?locationCluster
HAVING(COUNT(?observation) > 1) # Filter for extended time (e.g., >10 observations)
ORDER BY DESC(?durationCount)
```

```
LIMIT 10
```

Listing 50 SPARQL Query for Question 50 label

CQ51: *Will the elephants be arriving at DGFC soon?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephant ?currentLat ?currentLong ?speed ?direction ?distanceToDGFC ?date  ?time
{
  # Elephant observations
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude ?currentLat;
               pos:longitude ?currentLong;
               foo:speed ?speed;
               foo:direction ?direction ;
               foo:localDate   ?date;
               foo:localTime   ?time.
  # Coordinates of DGFC
  BIND(5.41382 AS ?dgfcLat)
  BIND(118.03771 AS ?dgfcLong)
  # Calculate distance to DGFC using a simplified formula
  BIND(
    SQRT(
      POW((?currentLat - ?dgfcLat) * 111.32, 2) + POW((?currentLong - ?dgfcLong)
      * 111.32, 2)
    ) AS ?distanceToDGFC
  )
  # Check if the elephants are moving toward DGFC
  FILTER (?distanceToDGFC < 1) # Threshold distance (e.g., within 1 km)
  FILTER (?speed > 0) # Moving elephants only
}
ORDER BY ?distanceToDGFC
```

Listing 51 SPARQL Query for Question 51 label

CQ52: *How many satellites did the collar detect? (COV=0, speed=0)*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
```

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephant (COUNT(?satellite) AS ?satelliteCount)
{
  # Observations with specific conditions
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               foo:speed ?speed;
               foo:cov ?COV;

  # Filters for COV=0 and speed=0
  FILTER (?COV = 0)
  FILTER (?speed = 0)
}
GROUP BY ?elephant
```

Listing 52 SPARQL Query for Question 52 label

CQ53 *Which elephants are close to the river today?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?elephant ?lat ?long ?distanceToRiver
WHERE {
  # Elephant observations
  ?observation    a                     foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude             ?lat;
               pos:longitude            ?long;
               foo:localDateTime        ?observationTime.
  # Today's␣date␣(filtering␣based␣on␣the␣current␣date)
␣␣FILTER␣(STRSTARTS(STR(?observationTime),␣STR(SUBSTR(STR(NOW()),␣1,␣10))))
␣␣#␣Coordinates␣of␣the␣river␣(approximation␣for␣the␣query)
␣␣BIND(5.5␣AS␣?riverLat)
␣␣BIND(118.0␣AS␣?riverLong)
␣␣#␣Calculate␣distance␣to␣the␣river␣using␣a␣simplified␣formula
␣␣BIND(
␣␣␣␣SQRT(
␣␣␣␣␣␣POW((?lat␣-␣?riverLat)␣*␣111.32,␣2)␣+␣POW((?long␣-␣?riverLong)␣*␣111.32,␣2)
␣␣␣␣)␣AS␣?distanceToRiver
␣␣)
␣␣#␣Filter␣for␣elephants␣close␣to␣the␣river␣(e.g.,␣within␣1␣km)
␣␣FILTER␣(?distanceToRiver␣<=␣1)
```

```
}
ORDER␣BY␣?elephant
```

Listing 53 SPARQL Query for Question 53 label

CQ54: *Which elephants are close to oil plantations?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
# For this query, the oil palm planation has been modelled inside FOO
# (w3id.org/def/foo#) on 20 November2024.
SELECT DISTINCT ?elephant ?lat ?long ?plantationLat ?plantationLong ?distanceToPlantation
{
  ?observation   a                      foo:gPSObservation;
                 foo:hasFeatureOfInterest ?elephant;
                 pos:latitude             ?lat;
                 pos:longitude            ?long.
  # Oil Palm Plantation locations
  foo:plantation   a                   foo:OilPalmPlantation ;
                pos:latitude           ?plantationLat;
                pos:longitude          ?plantationLong.
  # Calculate distance to oil palm plantations using a simplified formula
  BIND(
    SQRT(
      POW((?lat - ?plantationLat) * 111.32, 2) + POW((?long - ?plantationLong) * 111.32, 2)
    ) AS ?distanceToPlantation
  )
  # Filter for elephants close to plantations (e.g., within 5 km)
  FILTER (?distanceToPlantation <= 5)
}
ORDER BY ?elephant
```

Listing 54 SPARQL Query for Question 54 label

CQ55: *Which elephant roams near the Sabahmas site?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?elephant {
```

```
  # Elephant observations
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude ?lat;
               pos:longitude ?long .

#   # Approximate geographical filter for the Sabahmas site,
#  adjust coordinates as necessary
#   # This example uses a simple bounding box.
#   FILTER (?lat >= ?sabahmasLatMin && ?lat <= ?sabahmasLatMax &&
#           ?long >= ?sabahmasLongMin && ?long <= ?sabahmasLongMax)
}
```

Listing 55 SPARQL Query for Question 55 label

CQ56: *Which elephant roams near the small steep site?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?elephant {
  # Elephant observations
  ?observation a foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude ?lat; \\
               pos:longitude ?long .
  # # Placeholder coordinates for the ""small steep site"", replace with actual values
  # LET (?siteLat := 0.0)  # Replace with the latitude of the site
  # LET (?siteLong := 0.0)  # Replace with the longitude of the site
  # LET (?threshold := 0.01)  # Define a threshold for proximity, e.g., ~1km, adjust as needed

  # Filter for elephants near the site within the defined threshold
  # FILTER (ABS(?lat - ?siteLat) < ?threshold && ABS(?long - ?siteLong) < ?threshold)
}
```

Listing 56 SPARQL Query for Question 56 label

CQ57: *Which elephant is likely to visit Ribubonus, Kg. Kiabau, and Reka Halus 12ha?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

```
SELECT ?elephant (COUNT(?observation) AS ?visits)
{
   ?observation a                      foo:gPSObservation;
                foo:hasFeatureOfInterest ?elephant;
                pos:latitude            ?lat;
                pos:longitude           ?long .
  # Example filter for a location, assuming known coordinates or identifiers
  # Replace with actual conditions that define being ""near"" each location
#   FILTER (
#     # Conditions for Ribubonus
#     (ABS(?lat - ?ribubonusLat) < ?threshold && ABS(?long - ?ribubonusLong) < ?threshold)
#     OR
#     # Conditions for Kg. Kiabau
#     (ABS(?lat - ?kgKiabauLat) < ?threshold && ABS(?long - ?kgKiabauLong) < ?threshold)
#     OR
#     # Conditions for Reka Halus 12ha
#     (ABS(?lat - ?rekaHalusLat) < ?threshold && ABS(?long - ?rekaHalusLong) < ?threshold))
}
GROUP BY ?elephant
ORDER BY DESC(?visits)
```

Listing 57 SPARQL Query for Question 57 label

CQ58: *What locations could have snares?*

```
# Foo knowledge graph does not have snares location data.
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT ?location ?riskFactor{
  ?location dbo:hasRiskFactor ?riskFactor .
  ?riskFactor dbo:type dbr:Snare .}
```

Listing 58 SPARQL Query for Question 58 label

CQ59: *Is Elephant X sick, injured, or dead?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX elephant: <http://example.org/elephants#>
SELECT ?elephant ?speed ?cov (IF(?speed = 0 && ?cov = 0, "Dead",
IF(?speed > 0, "Alive", "Injured")) AS ?status)
```

```
{
   ?observation a                        foo:gPSObservation;
                foo:hasFeatureOfInterest ?elephant;
                pos:latitude             ?lat;
                pos:longitude            ?long;
                foo:speed                ?speed;
                foo:cov                  ?cov.}
ORDER BY DESC(?observation)
LIMIT 1
```

Listing 59 SPARQL Query for Question 59 label

CQ60: *Which elephant(s) are likely to conflict with human?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?elephant ?lat ?long (NOW() AS ?queryTime)
{
  ?observation a                        foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude             ?lat;
               pos:longitude            ?long;
               foo:localTime            ?time.
  # Filter for observations within the last month
  FILTER (?time > xsd:dateTime(NOW() - "P1M"^^xsd:duration))}
```

Listing 60 SPARQL Query for Question 60 label

CQ61: *What is the soil condition during certain times of the year?*

```
# Assume the area is "Danum_Valley_Conservation_Area"
#Since dates are not included in the data,
# only soil parameters are retrieved.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?observation a               foo:soilObservation ;
               foo:site        "Danum_Valley_Conservation_Area"^^ xsd:string ;
               foo:horizon     ?horizon ;
               foo:landUse     ?landUse ;
```

```
        foo:clay        ?Clay  ;
        foo:silt        ?Silt  ;
        foo:soilPH      ?soilPH ;
        foo:totalC      ?totalC ;
        foo:totalN      ?totalN  ;
        foo:cNRatio     ?CNRatio ;
        foo:totalP      ?totalP . }
```

Listing 61 SPARQL Query for Question 61 label

CQ62: *What types of soil are available throughout the year? Dry, muddy, swamps.*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT  ?soilType {
  ?observation a           foo:soilObservation ;
            foo:horizon    ?soilType.}
#soil's horizon resulted "Organic" "Mineral"
```

Listing 62 SPARQL Query for Question 62 label

CQ63: *Where are the locations of the type of soil that elephants prefer? (e.g., in the forest, near the river, open spaces, fields, and grass areas)*

```
# FOO can't answer this question, need more data.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
SELECT DISTINCT  ?soilType {
SELECT ?location ?soilType ?elephantDensity
WHERE {
  # Step 1: Find elephant observations and their proximity
  {
    SELECT ?location (COUNT(?elephant) AS ?elephantDensity)
    WHERE {
      # Elephant observations with coordinates
      ?elephant  a           foo:gPSObservation ;
                 foo:latitude  ?lat ;
                 foo:longitude ?long .
```

```
#␣Relate␣these␣elephants␣to␣a␣location␣(using␣a␣predefined␣spatial␣relationship)
␣␣␣␣␣␣?location␣a␣foo:Location␣;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣foo:hasBoundary␣?boundary␣.
␣␣␣␣␣␣FILTER(geof:sfWithin(?elephant,␣?boundary))

␣␣␣␣␣␣#␣Optional:␣filter␣for␣proximity␣between␣individual␣elephants
␣␣␣␣␣␣?otherElephant␣a␣␣␣␣␣␣␣␣␣␣␣foo:gPSObservation␣␣;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣foo:latitude␣␣␣␣?lat2␣;
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣foo:longitude␣␣␣?long2␣.
␣␣␣␣FILTER(bif:st_distance(bif:st_point(?lat,␣?long),␣bif:st_point(?lat2,␣?long2))
␣␣␣␣<␣1000)␣#␣within␣1␣km}
␣␣␣␣GROUP␣BY␣?location
␣␣␣␣HAVING␣(COUNT(?elephant)␣>␣1)␣#␣Ensure␣there␣is␣more␣than␣one␣elephant}
␣␣#␣Step␣2:␣Link␣the␣locations␣to␣soil␣types
␣␣?location␣foo:hasSoilType␣?soilType␣.
␣␣#␣Optional:␣Filter␣for␣soil␣types␣elephants␣prefer
␣␣FILTER␣(?soilType␣IN␣("forest_soil",␣"riverbank_soil",␣"open_space_soil",
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣"field_soil",␣"grassland_soil"))}
ORDER␣BY␣DESC(?elephantDensity)}
```

Listing 63 SPARQL Query for Question 63 label

CQ64: *What is the mineral content (salt and others) in a particular location?*

```
# Let a particular location be the oil palm plantation
# Soil pH: Determines acidity or alkalinity, influencing nutrient
availability and microbial activity.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?Observation ?soilPH ?acidityLevel ?lat ?long
WHERE {
  ?Observation a            foo:soilObservation ;
               foo:horizon "Mineral"^^xsd:string ;
               foo:soilPH   ?soilPH .
  # Specifying the location; replace with actual plantation coordinates
  foo:plantation pos:latitude ?lat ;
                 pos:longitude ?long .
  # Determine soil acidity or alkalinity
  BIND(
    IF(?soilPH < 7.0, "Acidic", "Alkaline") AS ?acidityLevel)}
```

Listing 64 SPARQL Query for Question 64 label

CQ65: *Is there any metal in the soil in that area?*

```
#FOO does not have metal data.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT * {
  ?Observation a                          foo:soilObservation ;
               foo:containsMetal          ?metal ;
               foo:soilMetalConcentration ?metalConcentration ;
               foo:horizon "Mineral"^^xsd:string .

  ?metal      foo:metalType                ?metalType .
  # Specifying the location; replace with actual plantation coordinates
  foo:plantation pos:latitude             ?lat ;
                 pos:longitude            ?long .

  # Optional: Filter for significant metal concentration (adjust threshold as needed)
  FILTER(?metalConcentration > 0)}
```

Listing 65 SPARQL Query for Question 65 label

CQ66: *What are the chemicals, agrochemical concentrations in the soil of a certain area?*

```
# To relate soil characteristics like totalC, totalN, soil pH, organicP, silt, clay, and
# sand to chemicals and agrochemical concentrations, we can integrate these components into
# a single SPARQL query. This query allows you to analyze the relationship
between agrochemical
# presence and soil properties at a specific location.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observation with components and agrochemical data
  ?Observation    a         foo:soilObservation ;
                 foo:totalC ?totalC ;
                 foo:totalN ?totalN ;
                 foo:soilPH ?soilPH ;
                 foo:silt    ?silt ;
                 foo:clay    ?clay ;
                 foo:sand    ?sand .
```

```
#   Specifying the location; replace with actual coordinates
  foo:plantation pos:latitude ?lat ;
                 pos:longitude ?long .
#   Optional Filters for meaningful thresholds
  FILTER(?soilPH >= 5.5 && ?soilPH <= 7.5)   # Neutral pH range
  FILTER(?totalC > 2)                        # Significant carbon content
}
```

Listing 66 SPARQL Query for Question 66 label

CQ67: *Does the soil in location X contain disease pathogens?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
# Soil observation with components and agrochemical data
  ?Observation a            foo:soilObservation ;
                 foo:totalC ?totalC ;
                 foo:totalN ?totalN ;
                 foo:soilPH ?soilPH ;
                 foo:silt    ?silt ;
                 foo:clay    ?clay ;
                 foo:sand    ?sand .
#   Specifying the location; replace with actual coordinates
  foo:plantation pos:latitude ?lat ;
                 pos:longitude ?long .
  # Optional Filters for conditions potentially supporting pathogens
  FILTER(?soilPH < 6.5 || ?soilPH > 7.5) # Favorable pH for pathogens
  (e.g., acidic or alkaline soils)
  FILTER(?totalC > 2) # High organic content, which can support pathogen growth
  FILTER(?silt + ?clay > 50) # Fine-textured soils with high water retention}
```

Listing 67 SPARQL Query for Question 67 label

CQ68: *Which area needs pesticide spraying?*

```
# To determine which area needs pesticide spraying,
# you would typically assess the soil and environmental
# conditions that favor pests or pathogens, or analyze data
# indicating pest infestations. If we do not have direct pest data,
we can infer risk areas based on environmental
```

```
#  factors (e.g., soil properties, pathogen risks, crop types).
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations and location data
  ?observation a              foo:soilObservation ;
               foo:soilPH     ?soilPH ;
               foo:inorganicP ?organicP ;
               foo:silt       ?silt ;
               foo:clay       ?clay ;
               foo:sand       ?sand ;
               foo:site       ?site .
  # Inference of risk level based on soil properties and environmental factors
  BIND(
    IF(?soilPH < 6.5 || ?soilPH > 7.5 || ?organicP > 1.5 || (?silt + ?clay > 50),
    "High_Risk", "Low_Risk") AS ?riskLevel)
  # Filter for high-risk areas that may need pesticide spraying
  FILTER(?riskLevel = "High_Risk")}
```

Listing 68 SPARQL Query for Question 68 label

CQ69: *What is the soil moisture level in a specific location?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations with moisture and clay content
  ?observation a          foo:soilObservation ;
               foo:clay   ?clay ;
               foo:site   ?site.
#   # Specify the target location; replace with actual coordinates
#   FILTER (?lat = SPECIFIC_LATITUDE^^xsd:decimal && ?long =
SPECIFIC_LONGITUDE^^xsd:decimal)}
```

Listing 69 SPARQL Query for Question 69 label

CQ70: *What is the presence of minerals in the soil?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *{
  ?observation a           foo:soilObservation;
               foo:site    ?site ;
               foo:horizon "Mineral".}
```

Listing 70 SPARQL Query for Question 70 label

CQ71: *Are there signs of heavy metal in the soil?*

```
#FOO does not have metal data.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT *  {
  ?Observation a                          foo:soilObservation ;
               foo:containsMetal          ?metal ;
               foo:soilMetalConcentration ?metalConcentration ;
               foo:horizon "Mineral"^^xsd:string .}
```

Listing 71 SPARQL Query for Question 71 label

CQ72: *Where are the salt licks located?*

```
# Danum Valley Conservation Area has saltlicks
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT *  {
  ?Observation a            foo:soilObservation ;
               foo:site     "Danum Valley Conservation Area";
               foo:horizon ?horizon .}
```

Listing 72 SPARQL Query for Question 72 label

CQ73: *What are the mineral and salt concentrations in the soil that indicate the presence of salt licks in a particular location?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations and associated data
  ?observation a          foo:soilObservation ;
               foo:site   ?site ;
               foo:totalP ?Phosphours ;
               foo:totalC ?Calcium .
  # Filter for high mineral concentrations
  (indicative of salt licks or significant deposits)
 FILTER (
    (?Calcium >= 300) ||
    (?Phosphours >= 100))}
```

Listing 73 SPARQL Query for Question 73 label

CQ74: *What is the pH level in the soil?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observation with components and agrochemical data
  ?Observation a            foo:soilObservation ;
                foo:soilPH ?soilPH. }
```

Listing 74 SPARQL Query for Question 74 label

CQ75: *What is the temperature reading from the soil sensor?*

```
# FOO soil data do not  contain temperature. GPS data do.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
?Observation a               foo:soilObservation ;
             foo:temperature  ?temperature .}
```

Listing 75 SPARQL Query for Question 75 label

CQ76: *What is the soil moisture in a certain location?*

```
# FOO does not contain soil moisture
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?observation a foo:soilObservation ;
               ?p ?o ;
               ?p1 ?o2.  }
```

Listing 76 SPARQL Query for Question 76 label

CQ77: *Is the soil in this area healthy for animals?*

```
# To determine whether the soil in a specific area is healthy for animals,
# the query should assess soil health metrics (e.g., pH) that influence animal health.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT *{
  # Soil observations
  ?observation a         foo:soilObservation ;
               foo:site  ?site ;
               foo:soilPH ?soilPH ;
               foo:totalP ?totalP ;
               foo:totalC ?totalC .
  # Soil health assessment
  BIND(
    IF(?soilPH >= 6.0 && ?soilPH <= 7.5 &&
       ?totalP > 50 && ?totalC > 2, "Healthy", "Unhealthy") AS ?healthStatus)}
```

Listing 77 SPARQL Query for Question 77 label

CQ78: *Is the soil fertile in this area?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
```

```
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT *
{# Soil observations
  ?observation a         foo:soilObservation ;
               foo:site ?site ;
               foo:soilPH ?soilPH ;
               foo:totalN ?totalN ;
               foo:totalP ?totalP ;
               foo:totalC ?totalC ;
               foo:silt ?silt ;
               foo:clay ?clay ;
               foo:sand ?sand .
  # Soil fertility assessment
  BIND(
    IF(
      (?soilPH >= 6.0 && ?soilPH <= 7.5) &&  # Optimal pH range
      (?totalN > 0.2) &&                     # Minimum nitrogen content
      (?totalP > 50) &&                      # Minimum phosphorus content
      (?totalC > 2) &&                       # Adequate organic carbon
      (?silt + ?clay + ?sand = 100),         # Ensure valid soil texture percentages
      "Fertile",
      "Infertile"
    ) AS ?fertilityStatus
  )
}
```

Listing 78 SPARQL Query for Question 78 label

CQ79: *What is the moisture rate of the soil in this area (i.e., provide geolocation)?*

```
# FOO needs location and soil moisture data
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT *{
  # Soil observations with moisture and location data
  ?observation a               foo:soilObservation ;
               foo:soilMoisture ?soilMoisture ;
               foo:site         ?Site ;
               pos:latitude     ?latitude ;
               pos:longitude    ?longitude .
```

```
  # Filter for a specific geolocation; replace with actual coordinates
  FILTER(?latitude = SPECIFIC_LATITUDE^^xsd:decimal && ?longitude
  = SPECIFIC_LONGITUDE^^xsd:decimal)}
```

Listing 79 SPARQL Query for Question 79 label

CQ80: *Where to plant crops for elephants (i.e., soil moisture rates)?*

```
# FOO needs location and soil moisture data
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT * {
  # Soil observations with location and moisture data
  ?Observation a          foo:soilObservation ;
               foo:soilPH ?soilPH ;
               foo:silt  ?silt ;
               foo:clay  ?clay ;
               foo:sand  ?sand ;
               foo:site  ?site .
  # Filter for soil pH suitable for crops (e.g., 6.0-7.5)
  FILTER(?soilPH >= 6.0 && ?soilPH <= 7.5)
  # Ensure valid soil texture percentages (silt + clay + sand = 100)
  FILTER((?silt + ?clay + ?sand) = 100)}
```

Listing 80 SPARQL Query for Question 80 label

CQ81: *Could planting in safer areas (healthy soil) influence animal movements?*

```
# Hypothesis:
# Planting in areas with healthy soil
(based on pH, organic content, nutrients, and absence of toxins)
encourages animal movement towards those areas due to:
# Availability of Nutritious Forage:
Plants grown in healthy soil are more palatable and nutrient-rich,
which attracts herbivores like elephants.
#Proximity to Resources: Healthy soil retains water better,
fostering consistent vegetation even during dry seasons,
influencing animal migration or foraging behavior.
#Avoidance of Toxicity: Healthy soil reduces risks of harmful elements
(e.g., heavy metals) affecting vegetation quality and animal health.
```

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # elephant movement observations
  ?movement a                        foo:gPSObservation ;
            foo:hasFeatureOfInterest ?elephant;
            foo:madeBySensor         ?sensor ;
            foo:localTime            ?movementTime ;
            pos:latitude             ?latitude ;
            pos:longitude            ?longitude .
  # Soil health data for planting sites
  ?soil a                           foo:soilObservation ;
          foo:soilPH                ?soilPH ;
          foo:totalC                ?totalC ;
          foo:totalP                ?totalP .
  # Filter for healthy soil characteristics
  FILTER(?soilPH >= 6.0 && ?soilPH <= 7.5)  # Optimal soil pH
  FILTER(?totalC > 2)                       # Sufficient organic carbon
  FILTER(?totalP > 50)                      # Adequate phosphorus
}
```

Listing 81 SPARQL Query for Question 81 label

CQ82: *Could we predict crop yield based on soil data?*

```
# FOO needs crop data.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations
  ?observation a             foo:soilObservation ;
               foo:site      ?site ;
               foo:soilPH    ?soilPH ;
               foo:totalN    ?totalN ;
               foo:totalP    ?totalP ;
               foo:totalK    ?totalK ;
               foo:totalC    ?totalC ;
               foo:silt      ?silt ;
               foo:clay      ?clay ;
```

```
              foo:sand      ?sand .
  # Crop data
  ?plantingSite a            foo:PlantingSite ;
                foo:cropType ?cropType ;
                pos:latitude ?latitude ;
                pos:longitude ?longitude .
  # Soil fertility assessment
  BIND(
    IF(
      (?soilPH >= 6.0 && ?soilPH <= 7.5) &&  # Optimal soil pH
      (?totalN > 0.2) &&                    # Sufficient nitrogen
      (?totalP > 50) &&                     # Sufficient phosphorus
      (?totalC > 2) &&                      # Sufficient organic carbon
      (?silt + ?clay + ?sand = 100),        # Valid soil texture
      "Fertile",
      "Infertile"
    ) AS ?fertilityStatus
  )
  # Predicted crop yield (simplified prediction logic for demonstration)
  BIND(
    IF(
      ?fertilityStatus = "Fertile" && ?soilMoisture >= 20 && ?soilMoisture <= 35,
      "High Yield",
      "Low Yield"
    ) AS ?predictedYield
  )
}
```

Listing 82 SPARQL Query for Question 82 label

CQ83: *What soil metrics help us predict floods?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT * {
  # Soil observations with key metrics
  ?observation a             foo:soilObservation ;
               foo:site      ?site ;
               foo:soilPH    ?soilPH ;
               foo:totalN    ?totalN ;
               foo:totalP    ?totalP ;
```

```
                foo:totalC    ?totalC ;
                foo:silt      ?silt ;
                foo:clay      ?clay ;
                foo:sand      ?sand .
  # Calculating flood risk based on soil texture and organic carbon
  BIND(
    IF(
      (?clay > 40 && ?sand < 20) ||  # High clay content indicates poor drainage
      (?silt > 40) ||               # Silt-heavy soils can lead to surface runoff
      (?totalC < 2),                # Low organic carbon reduces water retention
      "High Risk",
      "Low Risk"
    ) AS ?floodRisk
  )
}
```

Listing 83 SPARQL Query for Question 83 label

CQ84: *What are the metrics of healthy soil with less/no chemical pollution from oil palm plantations?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations
  ?observation a            foo:soilObservation ;
                foo:site    ?Site ;
                foo:soilPH  ?soilPH ;
                foo:totalC  ?totalC ;
                foo:totalN  ?totalN ;
                foo:totalP  ?totalP ;
                foo:silt    ?silt ;
                foo:clay    ?clay ;
                foo:sand    ?sand .
# Oil Palm Plantation locations
  foo:plantation   a        foo:OilPalmPlantation ;
               pos:latitude  ?plantationLat;
               pos:longitude ?plantationLong.
  # Soil health assessment
  BIND(
    IF(
      (?soilPH >= 6.0 && ?soilPH <= 7.5) &&       # Healthy pH range
```

```
      (?totalC > 2) &&           # Sufficient organic carbon
      (?totalN > 0.2) &&         # Adequate nitrogen
      (?totalP > 50),            # Adequate phosphorus
      "Healthy",                 # Health classification
      "Polluted"
    ) AS ?healthStatus
  )
}
```

Listing 84 SPARQL Query for Question 84 label

CQ85: *Why do elephants not like to walk on wet soil (movement prediction)?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations with elevation and soil metrics
  ?soilObservation    a      foo:soilObservation ;
                      foo:site      ?Site ;
                      foo:clay      ?clay .
    # GPS elephant movements and elevations
    ?ElephantObservation a   foo:gPSObservation ;
                      foo:altitude  ?elevation ;
                      pos:latitude  ?latitude ;
                      pos:longitude ?longitude .
#   # Predict movement probability based on elevation and soil metrics
  BIND(
    IF(
      (?clay > 40 || ?elevation < 50), # Wet, clay-heavy, or low elevation
      "Low",             # Low probability of movement
      "High"             # High probability of movement
    ) AS ?movementProbability
  )
}
```

Listing 85 SPARQL Query for Question 85 label

CQ86: *What are the chemical levels of the soil in Protected Area 1?*

```
# Let Protected Area 1 be "Maliau_Basin_Conservation_Area"
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations in Protected Area 1
  ?observation a            foo:soilObservation ;
               foo:site     ?site;
               foo:totalN   ?totalN ;
               foo:totalP   ?totalP ;
               foo:totalC   ?totalC .
  # Filter for Protected Area 1 (replace with actual identifier or coordinates)
  FILTER(?site = "Maliau_Basin_Conservation_Area")
}
```

Listing 86 SPARQL Query for Question 86 label

CQ87: *What are the soil nutrient levels?*

```
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX foo: <https://w3id.org/def/foo#>
SELECT *
{?observation   a            foo:soilObservation ;
                foo:site     ?site   ;
                foo:totalN   ?totalN ;
                foo:totalP   ?totalP ;
                foo:totalC   ?totalC .
FILTER (STRSTARTS(STR(?site), "D")) # search for sites }
```

Listing 87 SPARQL Query for Question 87 label

CQ88: *What is the effect of moisture on nutrients and oxygen levels?*

```
# Let Protected Area 1 be "Maliau_Basin_Conservation_Area"
# neagtive correlation of -0.999
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT
```

```
    (SUM(?moistureLevel * ?oxygenLevel) AS ?sumXY)
    (SUM(?moistureLevel) AS ?sumX)
    (SUM(?oxygenLevel) AS ?sumY)
    (SUM(?moistureLevel * ?moistureLevel) AS ?sumX2)
    (SUM(?oxygenLevel * ?oxygenLevel) AS ?sumY2)
    (COUNT(?moistureLevel) AS ?n)
    ((?n * ?sumXY - ?sumX * ?sumY) /
     SQRT((?n * ?sumX2 - ?sumX * ?sumX) * (?n * ?sumY2 - ?sumY * ?sumY)) AS ?correlation)
 {
  ?observation a             foo:soilObservation ;
               foo:site      ?site ;
               foo:soilPH    ?soilPH ;
               foo:totalC    ?totalC ;
               foo:totalN    ?totalN ;
               foo:totalP    ?totalP .
  # Calculate moisture level
  BIND((?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7)) * 0.2)
    AS ?moistureLevel)
  # Calculate oxygen level
  BIND((20 - (?moistureLevel * 0.6) + (?totalC * 0.3) - (ABS(?soilPH - 7) * 0.4))
    AS ?oxygenLevel)}
```

Listing 88 SPARQL Query for Question 88 label

CQ89: *What is the ideal soil moisture rate for an elephant to give birth?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT
       ?site
       ?soilPH
       ?totalC
       ?totalN
       ?totalP
       ((?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7)) * 0.2)
       AS ?soilMoisture)
{
  # Soil observations
  ?observation a            foo:soilObservation ;
               foo:site     ?site ;
               foo:soilPH   ?soilPH ;
               foo:totalC   ?totalC ;
```

```
              foo:totalN   ?totalN ;
              foo:totalP   ?totalP .

  # Filter for ideal soil moisture range suitable for elephant births (20-35 is the range)
  FILTER(
    (?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7)) * 0.2) >= 20
    &&
    (?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7)) * 0.2) <= 35
  )
}
```

Listing 89 SPARQL Query for Question 89 label

CQ90: *What are the soil conditions in areas that have elephant grass?*

```
# FOO needs to bind them with common geo-locations.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  # Soil observations with elevation and soil metrics
  ?soilObservation a      foo:soilObservation ;
              foo:site   ?Site ;
              foo:clay  ?clay .
 # areas with elephant grass
 ?vegObservation  a                 foo:treeObservation  ;
               foo:date             ?date ;
               foo:id               ?id ;
               foo:treeID           ?treeID  ;
               foo:treeIndividualNo ?treeindividual ;
               foo:treeNLianas      ?treeN;
               foo:treeNotes        ?treeNotes. }
```

Listing 90 SPARQL Query for Question 90 label

CQ91: *How to conserve suitable soils for the elephants to have food in the future (e.g., reduce the use of fertilizer)?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

```
SELECT DISTINCT
       ?site
       ?soilPH
       ?totalC
       ?totalN
       ?totalP
       ((?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7)) * 0.2)
       AS ?soilFertility)
WHERE {
  # Soil observations
  ?observation a            foo:soilObservation ;
               foo:site     ?site ;
               foo:soilPH   ?soilPH ;
               foo:totalC   ?totalC ;
               foo:totalN   ?totalN ;
               foo:totalP   ?totalP .
  # Filter for high-fertility soils
  FILTER((?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7))
  * 0.2) >= 25)}
ORDER BY DESC((?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7)) * 0.2))
```

Listing 91 SPARQL Query for Question 91 label

CQ92: *What soil moisture do elephants spend most time on?*

```
# Based on palm leaves (food) and logged urban areas (oil palm plantation).
Query took 300959ms to run.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT
       ?elephant
       ?lat
       ?long
       ?plantationLat
       ?plantationLong
       ?distanceToPlantation
       ?site
       ?soilPH
       ?totalC
       ?totalN
       ?totalP
       ((?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 +
```

```
      (7 - ABS(?soilPH - 7)) * 0.2) AS ?soilMoisture)
WHERE {
  # Elephant GPS observations
  ?observation a                        foo:gPSObservation;
               foo:hasFeatureOfInterest  foo:jasmin;
               pos:latitude             ?lat;
               pos:longitude            ?long.

  # Oil Palm Plantation locations
  foo:plantation a foo:OilPalmPlantation;
                 pos:latitude ?plantationLat;
                 pos:longitude ?plantationLong.
  # Calculate approximate distance between elephant and plantation
  BIND(
    111.32 * SQRT(
      POW((?lat - ?plantationLat), 2) + POW((?long - ?plantationLong), 2)
    ) AS ?distanceToPlantation
  )
  # Soil observations
  ?soilObservation a             foo:soilObservation;
                   foo:site      ?site;
                   foo:soilPH    ?soilPH;
                   foo:totalC    ?totalC;
                   foo:totalN    ?totalN;
                   foo:totalP    ?totalP.
  # Link soil observations to elephant observations
  FILTER(
    ?distanceToPlantation < 5 # Assuming 5 km radius for proximity to plantation
  )
}
ORDER BY ASC(?distanceToPlantation) DESC(?soilMoisture)
```

Listing 92 SPARQL Query for Question 92 label

CQ93: *What do elephants eat? Provide one example for each area with elephant grass (Napier), other grass, bark, palm shoots, young leaves, trunks, soft plants, bananas?*

```
 # FOO needs elephant food types data (Future Work)
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT
       ?elephant
```

```
      ?lat
      ?long
      ?plantationLat
      ?plantationLong
      ?distanceToPlantation
      ?site
      ?soilPH
      ?totalC
      ?totalN
      ?totalP
      ?foodType
      (SAMPLE(?foodExample) AS ?example)
      ((?totalC * 0.8 + ?totalN * 0.5 + ?totalP * 0.3 + (7 - ABS(?soilPH - 7)) * 0.2)
      AS ?soilMoisture)
{
  ?observation a                        foo:gPSObservation;
               foo:hasFeatureOfInterest ?elephant;
               pos:latitude             ?lat;
               pos:longitude            ?long.

  # Oil Palm Plantation locations
  foo:plantation a foo:OilPalmPlantation;
                 pos:latitude ?plantationLat;
                 pos:longitude ?plantationLong.
  # Calculate approximate distance between elephant and plantation
  BIND(
    111.32 * SQRT(
      POW((?lat - ?plantationLat), 2) + POW((?long - ?plantationLong), 2)
    ) AS ?distanceToPlantation )
  # Soil observations
  ?soilObservation a              foo:soilObservation;
                   foo:site       ?site;
                   foo:soilPH     ?soilPH;
                   foo:totalC     ?totalC;
                   foo:totalN     ?totalN;
                   foo:totalP     ?totalP.
  # Feeding observations
  ?feedingObservation a                     foo:FeedingObservation;
                   foo:hasFeatureOfInterest ?elephant;
                      foo:consumes          ?foodExample.
  # Link food example to type
  ?foodExample foo:foodType ?foodType.
  # Filter for specific food types
  FILTER(?foodType IN (
```

```
    "elephant␣grass",
    "other␣grass",
    "bark",
    "palm␣shoots",
    "young␣leaves",
    "trunks",
    "soft␣plants",
    "bananas" ))
  # Link soil observations to elephant observations
  FILTER(?distanceToPlantation < 10)  # Assuming 10 km radius for proximity to plantation}
GROUP BY ?elephant ?lat ?long ?plantationLat ?plantationLong ?distanceToPlantation
?site ?soilPH ?totalC ?totalN ?totalP ?foodType
ORDER BY ASC(?distanceToPlantation) DESC(?soilMoisture)
```

Listing 93 SPARQL Query for Question 93 label

CQ94: *Where do bamboo shoots grow?*

```
# FOO requires :Bamboo Plantation informantion
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?bambooPlantation  a     foo:BambooPlantation ;
           pos:latitude    ?latitude ;
           pos:longitude   ?longitude .}
```

Listing 94 SPARQL Query for Question 94 label

CQ95: *Where could we find areas with the inner trunk of oil palms?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT * {
  ?oilpalm   a            foo:OilPalmPlantation;
           pos:latitude   ?latitude ;
           pos:longitude  ?longitude .}
```

Listing 95 SPARQL Query for Question 95 label

CQ96: *Where could we find areas with broad leaves?*

```
# In Danum Valley Conservation Area and Maliau Basin
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
{
  ?area  a            foo:treeObservation ;
         foo:siteName  ?site;

  FILTER(regex(?site, "Danum|Maliau Basin", "i"))
}
```

Listing 96 SPARQL Query for Question 96 label

CQ97: *Where could we find areas with vines?*

```
# In Kinabatangan River Basin
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
{
  ?area a            foo:treeObservation ;
        foo:siteName  ?site;

  FILTER(regex(?site, "Kinabatangan River Basin", "i"))
}
```

Listing 97 SPARQL Query for Question 97 label

CQ98: *How can vegetation and site habitat information help understand the future patterns/locations of elephants?*

```
PREFIX foo: <http://w3id.org/def/foo#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX unit: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
```

```
SELECT * {
  ?s    a              foo:gPSObservation;
        foo:longitude  ?long;
        foo:latitude   ?lat.
BIND(STRDT(CONCAT('POINT(', STR(?long), '␣', STR(?lat), ')'), geo:wktLiteral) AS ?geom)
BIND(STRDT('POINT(118.33␣␣5.42)', geo:wktLiteral) AS ?targetGeom)
# FILTER(geof:distance(?geom, ?targetGeom, unit:Meter) < 10000)}
```

Listing 98 SPARQL Query for Question 98 label

CQ99: *Do elephants drink lots of water?*

```
# FOO needs water source data
PREFIX foo: <http://w3id.org/def/foo#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX unit: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
SELECT ?elephant ?name ?waterConsumption ?waterSource ?distance ?latitude ?longitude
{
  ?elephant a                         foo:gPSObservation ;
            foo:name                  ?name ;
            foo:dailyWaterConsumption ?waterConsumption ;
            pos:latitude              ?latitude ;
            pos:longitude             ?longitude ;
            foo:nearbyWaterSource     ?waterSource .
  ?waterSource a foo:WaterSource ;
               pos:latitude           ?waterLat ;
               pos:longitude          ?waterLong .
  BIND(geof:distance(?latitude, ?longitude, ?waterLat, ?waterLong) AS ?distance)
  FILTER(?distance < 5000)  # Only elephants within 5 km of a water source
  OPTIONAL {
    ?elephant foo:waterNeeds         ?waterNeeds .}
}
```

Listing 99 SPARQL Query for Question 99 label

CQ100: *Where do we find fruit farms in Lower Kinabatangan?*

```
# FOO can be expanded with agriculture data (fruits) near Kinabatangan River.
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
```

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX unit: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
SELECT ?farm ?farmName ?fruitType ?latitude ?longitude
{?farm a                  foo:FruitFarm ;
        foo:name          ?farmName ;
        foo:produces      ?fruitType ;
        pos:latitude      ?latitude ;
        pos:longitude     ?longitude ;
        geo:nearby foo:KinabatanganRiver .
?fruitType a foo:Fruit .
FILTER(regex(?farmName, "Lower␣Kinabatangan", "i"))}
```

Listing 100 SPARQL Query for Question 100 label

CQ101: *What areas have fewer trees?*

```
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
{?observation a               foo:treeObservation;
           foo:treeID         ?treeID ;
           foo:lianaDbhCm     ?lianaDbhCm ;
           foo:treeDbhCm      ?treeDbhCm ;
           foo:treeHeightM    ?treeHeightM ;
           foo:treeNLianas    ?treeNLianas ;
           foo:siteName       ?siteName ;
           foo:plotNo ?plotNo .
  FILTER(?treeNLianas < 2)  # Filter for trees with fewer than 2 lianas}
```

Listing 101 SPARQL Query for Question 101 label

CQ102: *What plant species should be conserved in the areas the elephants visit?*

```
# Vegetation data require geo-location information.
PREFIX foo: <http://w3id.org/def/foo#>
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT *
{# GPS observation identifies areas elephants visit
  ?observation a                       foo:gPSObservation;
```

```
              foo:hasFeatureOfInterest ?elephant;
              pos:latitude             ?lat;
              pos:longitude            ?long.
  # Tree observations in the same area
  ?treeobservation a                   foo:treeObservation;
              foo:treeID               ?treeID ;
              foo:lianaDbhCm           ?lianaDbhCm ;
              foo:treeDbhCm            ?treeDbhCm ;
              foo:treeHeightM          ?treeHeightM ;
              foo:treeNLianas          ?treeNLianas ;
              foo:siteName             ?siteName ;
              foo:plotNo               ?plotNo ;
              pos:latitude             ?lat;
              pos:longitude            ?long.}
ORDER BY ?lat ?long
```

Listing 102 SPARQL Query for Question 102 label

CQ103: *What plant species effected by deforestation?*

```
# FOO does not contain deforestation data.
PREFIX foo: <http://w3id.org/def/foo#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
SELECT DISTINCT ?plantSpecies ?deforestedArea ?impactLevel
WHERE {
  # Identify areas affected by deforestation
  ?area foo:hasDeforestationStatus "Deforested" ;
        foo:impactLevel ?impactLevel .
  # Link plant species to these areas
  ?area foo:containsPlant ?plant .
  ?plant foo:speciesName ?plantSpecies .
  # Optional: Include additional details like region or deforestation cause
  OPTIONAL {
    ?area foo:region ?deforestedArea .
  }
}
ORDER BY ?impactLevel ?plantSpecies
```

Listing 103 SPARQL Query for Question 103 label

CQ104: *Which plant species are cultivated by the Grow Borneo project?*

```
PREFIX foo: <http://w3id.org/def/foo#>
SELECT ?species
{
  ?species a foo:Tree ;
           foo:isPlantedIn foo:growBorneo .
}
```

Listing 104 SPARQL Query for Question 104 label

CQ105: *How many trees has the Grow Borneo project planted in the last five years?*

```
# Tree count data are not available in FOO as of 2024.
PREFIX foo: <http://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT (SUM(?count) AS ?totalTrees)
{
?plantingEvent foo:isPartOfProject foo:growBorneo ;
                 foo:year          ?year ;
                 foo:treeCount     ?count .
  FILTER(?year >= xsd:date("2019-01-01"))
}
```

Listing 105 SPARQL Query for Question 105 label

CQ105: *Semantic Rule Language (SWRL) rule for hazard alert*

```
INSERT {
  ?s   a <https://w3id.org/def/foo#gPSObservation>;
         <https://w3id.org/def/foo#hazard> ?hazard.
}
WHERE {
?s  a  <https://w3id.org/def/foo#gPSObservation>;
       <https://w3id.org/def/foo#localDate>   ?data ;
       <https://w3id.org/def/foo#localTime>   ?time ;
       <http://www.w3.org/2003/01/geo/wgs84_pos#latitude> ?lat;
       <http://www.w3.org/2003/01/geo/wgs84_pos#longitude> ?long.
  # Retrieve plantation details
  <https://w3id.org/def/foo#plantation> a <https://w3id.org/def/foo#OilPalmPlantation>;
                                        <http://www.w3.org/2003/01/geo/wgs84_pos#latitude> ?plantationL
                                        <http://www.w3.org/2003/01/geo/wgs84_pos#longitude> ?plantation
  # Convert coordinates to float (if stored as literals)
```

```
BIND(xsd:float(?lat) AS ?latitude)
BIND(xsd:float(?long) AS ?longitude)
BIND(xsd:float(?plantationLat) AS ?oilpalmLat)
BIND(xsd:float(?plantationLong) AS ?oilpalmLong)
# Calculate distance using the Haversine formula
BIND(6371 * 2 * ASIN(SQRT(
  POW(SIN((?latitude - ?oilpalmLat) * PI() / 180 / 2), 2) +
  COS(?oilpalmLat * PI() / 180) * COS(?latitude * PI() / 180) *
  POW(SIN((?longitude - ?oilpalmLong) * PI() / 180 / 2), 2)
)) AS ?distance)
# Determine poaching based on the calculated distance
BIND(IF(?distance <= 5, 1, 0) AS ?hazard. }
```

Listing 106 SWRL for hazard alert label

	Open Data		Sensor Data	
Natural Language Statements (NLSs) reflected in FOO	Soil	Veg	GPS	Image
NLS1 Tracking elephant locations so that the wildlife department can give warnings to local people about the arrival of elephants.			*	*
NLS2 Examples of areas with elephant grass (Nappier), other grasses, bark, palm shoots, young leaf trunks, soft plants, and bananas.		*		
NLS3 Focus on the area of Lower Kinabatangan and the 14 collared elephants living there.			*	
NLS 4 Collared elephants will not go to primary forest sites.		*	*	
NLS5 The datasets in this research could be used to generate predictions.			*	
NLS6 Elephants do not intend to cause damage. It may occur when their strong and huge bodies come in contact with things.			*	*
NLS 7 Nearly all wild pigs in the area of Kinabatangan died from influenza viruses.				
NLS 8 There was a famous story about the rhino who lost one leg from poaching. It survived on three legs for a long time.				*
NLS9 Female Asian elephants are tusk-less.			*	*
NL10 Male Asian elephants are more likely to explore human areas than females, attracted by food.			*	

Table 1 Natural language statements and what data set can fulfil the task.

NLS3: *NLS3 Focus on the area of Lower Kinabatangan and the 14 collared elephants living there.*

```
# 14 modelled elephants are Aqeela (Female), Liun (Female),
Jasmin (Female), Putut (Female), Puteri (Female),
# Ita (Female), Sejati (Male), Sandi (Female),
Kasih (Female), Gading (Male), Ratu (Female),
Koyah (Female), Girang (Female-poisoned), Sandy (Male-found dead).
PREFIX pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX foo: <https://w3id.org/def/foo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
{
  ?elephants  a  foo:ElephasMaximus;
              foo:location  ?location.
#  Filter to focus on Lower Kinabatangan
 FILTER regex(?location, "Lower Kinabatangan", "i")
}
```

Listing 107 SPARQL Query for Question 105 label

SWRL01: *SWRL Rule in Turtle format: Detect Poaching Observations Near Oil Palm Plantations*

```
@prefix foo: <https://w3id.org/def/foo#> .
@prefix swrl: <http://www.w3.org/2003/11/swrl#> .
@prefix swrlb: <http://www.w3.org/2003/11/swrlb#> .
```

```
@prefix pos: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

### Define the SWRL Rule ###
foo:NearPlantationRule a swrl:Imp ;
    swrl:body (
        [ a swrl:ClassAtom ;
          swrl:classPredicate foo:gPSObservation ;
          swrl:argument1 ?s
        ]
        [ a swrl:DatavaluedPropertyAtom ;
          swrl:propertyPredicate pos:latitude ;
          swrl:argument1 ?s ;
          swrl:argument2 ?lat
        ]
        [ a swrl:DatavaluedPropertyAtom ;
          swrl:propertyPredicate pos:longitude ;
          swrl:argument1 ?s ;
          swrl:argument2 ?long
        ]
        [ a swrl:ClassAtom ;
          swrl:classPredicate foo:OilPalmPlantation ;
          swrl:argument1 ?plantation
        ]
        [ a swrl:DatavaluedPropertyAtom ;
          swrl:propertyPredicate pos:latitude ;
          swrl:argument1 ?plantation ;
          swrl:argument2 ?plantationLat
        ]
        [ a swrl:DatavaluedPropertyAtom ;
          swrl:propertyPredicate pos:longitude ;
          swrl:argument1 ?plantation ;
          swrl:argument2 ?plantationLong
        ]
        [ a swrl:BuiltInAtom ;
          swrl:builtin swrlb:subtract ;
          swrl:arguments (?latDiff ?lat ?plantationLat)
        ]
        [ a swrl:BuiltInAtom ;
          swrl:builtin swrlb:subtract ;
          swrl:arguments (?longDiff ?long ?plantationLong)
        ]
        [ a swrl:BuiltInAtom ;
          swrl:builtin swrlb:multiply ;
```

```
  swrl:arguments (?latRadDiff ?latDiff 3.14159)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:divide ;
  swrl:arguments (?latRadDiff ?latRadDiff 180)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:multiply ;
  swrl:arguments (?longRadDiff ?longDiff 3.14159)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:divide ;
  swrl:arguments (?longRadDiff ?longRadDiff 180)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:sin ;
  swrl:arguments (?sinLatDiffHalf (/ ?latRadDiff 2))
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:sin ;
  swrl:arguments (?sinLongDiffHalf (/ ?longRadDiff 2))
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:pow ;
  swrl:arguments (?sinLatDiffHalfSq ?sinLatDiffHalf 2)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:pow ;
  swrl:arguments (?sinLongDiffHalfSq ?sinLongDiffHalf 2)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:cos ;
  swrl:arguments (?cosLat1 (/ ?lat 180 * 3.14159))
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:cos ;
  swrl:arguments (?cosLat2 (/ ?plantationLat 180 * 3.14159))
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:multiply ;
  swrl:arguments (?cosMult ?cosLat1 ?cosLat2)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:add ;
```

```
          swrl:arguments (?haversine ?sinLatDiffHalfSq ?cosMult)
        ]
        [ a swrl:BuiltInAtom ;
          swrl:builtin swrlb:sqrt ;
          swrl:arguments (?sqrtHaversine ?haversine)
        ]
        [ a swrl:BuiltInAtom ;
          swrl:builtin swrlb:asin ;
          swrl:arguments (?asinHaversine ?sqrtHaversine)
        ]
        [ a swrl:BuiltInAtom ;
          swrl:builtin swrlb:multiply ;
          swrl:arguments (?distance 6371 * 2 ?asinHaversine)
        ]
        [ a swrl:BuiltInAtom ;
          swrl:builtin swrlb:lessThanOrEqual ;
          swrl:arguments (?distance 5)
        ]
    ) ;
    swrl:head (
        [ a swrl:DatavaluedPropertyAtom ;
          swrl:propertyPredicate foo:poaching ;
          swrl:argument1 ?s ;
          swrl:argument2 true
        ]
    ) .
```

Listing 108 SWRL Rule label

SWRL02: *SWRL Rule: Identify GPS Observations Near Oil Palm Plantations as Hazard Areas*

```
@prefix foo: <https://w3id.org/def/foo#> .
@prefix swrl: <http://www.w3.org/2003/11/swrl#> .
@prefix swrlb: <http://www.w3.org/2003/11/swrlb#> .
@prefix pos: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

### Define the SWRL Rule ###
foo:GPSObservationToPlantationHazardRule a swrl:Imp ;
    swrl:body (
        [ a swrl:ClassAtom ;
          swrl:classPredicate foo:gPSObservation ;
          swrl:argument1 ?observation
```

```
]
[ a swrl:DatavaluedPropertyAtom ;
  swrl:propertyPredicate pos:latitude ;
  swrl:argument1 ?observation ;
  swrl:argument2 ?obsLat
]
[ a swrl:DatavaluedPropertyAtom ;
  swrl:propertyPredicate pos:longitude ;
  swrl:argument1 ?observation ;
  swrl:argument2 ?obsLong
]
[ a swrl:ClassAtom ;
  swrl:classPredicate foo:OilPalmPlantation ;
  swrl:argument1 ?plantation
]
[ a swrl:DatavaluedPropertyAtom ;
  swrl:propertyPredicate pos:latitude ;
  swrl:argument1 ?plantation ;
  swrl:argument2 ?plantationLat
]
[ a swrl:DatavaluedPropertyAtom ;
  swrl:propertyPredicate pos:longitude ;
  swrl:argument1 ?plantation ;
  swrl:argument2 ?plantationLong
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:subtract ;
  swrl:arguments (?latDiff ?obsLat ?plantationLat)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:subtract ;
  swrl:arguments (?longDiff ?obsLong ?plantationLong)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:pow ;
  swrl:arguments (?latDiffSq ?latDiff 2)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:pow ;
  swrl:arguments (?longDiffSq ?longDiff 2)
]
[ a swrl:BuiltInAtom ;
  swrl:builtin swrlb:add ;
  swrl:arguments (?geoDistSq ?latDiffSq ?longDiffSq)
```

```
    ]
    [ a swrl:BuiltInAtom ;
      swrl:builtin swrlb:sqrt ;
      swrl:arguments (?geoDistance ?geoDistSq)
    ]
    [ a swrl:BuiltInAtom ;
      swrl:builtin swrlb:lessThanOrEqual ;
      swrl:arguments (?geoDistance 5)
    ]
) ;
swrl:head (
    [ a swrl:ClassAtom ;
      swrl:classPredicate foo:HazardArea ;
      swrl:argument1 ?observation
    ]
) .
```

Listing 109 SWRL Rule label

.4 Appendix II: Forest Observatory Ontology (FOO)

```
@prefix cc: <http://creativecommons.org/ns#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix foo: <https://w3id.org/def/foo#> .
@prefix ns1: <http://data.bioontology.org/metadata/> .
@prefix ns2: <http://www.w3.org/2003/06/sw-vocab-status/ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix pos: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <https://schema.org/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix vann: <http://purl.org/vocab/vann/> .
@prefix wo: <https://www.bbc.co.uk/ontologies/wildlife-ontology#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

### Ontology Declaration ###
foo: a owl:Ontology ;
    cc:license <http://creativecommons.org/licenses/by-sa/4.0/> ;
    dc:abstract "The Forest Observatory Ontology (FOO) comprises a novel
    ontology that integrates wildlife data generated by sensors.
    FOO borrows classes and properties from SOSA and
    BBC wildlife ontology."@en ;
    dc:contributor [
        foaf:name "Professor Omer Rana"@en ;
        foaf:homepage <https://profiles.cardiff.ac.uk/staff/ranaof> ;
        schema:identifier <https://orcid.org/0000-0003-3597-2646>
    ],
    [
        foaf:name "Dr. Pablo Orozco-terWengel"@en ;
        foaf:homepage <profiles.cardiff.ac.uk/staff/orozco-terwengelpa> ;
        schema:identifier <orcid.org/0000-0002-7951-4148>
    ],
    [
        foaf:name "Professor Benoit Goossens"@en ;
        foaf:homepage <profiles.cardiff.ac.uk/staff/goossensbr> ;
        schema:identifier <https://orcid.org/0000-0003-2360-4643>
    ],
    [
        foaf:name "Dr. Charith Perera"@en ;
```

```
        foaf:homepage <profiles.cardiff.ac.uk/staff/pererac> ;
        schema:identifier <https://orcid.org/0000-0002-0190-3346>
    ] ;
    dc:creator [
        foaf:name "Naeima Hamed"@en ;
        foaf:homepage <cardiff.ac.uk/people/research-students/
        view/2501164-hamed-naeima> ;
        schema:identifier <orcid.org/0000-0002-2998-5056>
    ] ;
    dc:description "This ontology describes wildlife observations
    generated by sensors."@en ;
    dc:title "Forest Observatory Ontology (FOO)"@en ;
    dcterms:issued "2024-06-01"^^xsd:date ;
    dcterms:license <http://creativecommons.org/licenses/by-sa/4.0/> ;
    dcterms:publisher <https://ontoology.linkeddata.es/> ;
    vann:preferredNamespacePrefix "foo"@en ;
    vann:preferredNamespaceUri "https://w3id.org/def/foo#" ;
    owl:imports sosa: ;
    owl:versionIRI foo:V2.0 ;
    owl:versionInfo "BBC Wildlife Ontology Reused 26 June 2024" ;
    prov:generatedAtTime "2024-06-01T00:00:00+00:00"^^xsd:dateTime ;
    prov:wasAttributedTo <https://github.com/Naeima> ;
    prov:wasDerivedFrom <https://ontology.forest-observatory.org> ;
    schema:citation "Cite this vocabulary as: Hamed, N., Rana, O.,
    Goossens, B., Orozco-terWengel, P., Perera, C. (2023).
    FOO: An Upper-Level Ontology for the Forest Observatory.
    In: Pesquita, C., et al. The Semantic Web: ESWC 2023
    Satellite Events. ESWC 2023. Lecture Notes in Computer Science,
    vol 13998. Springer, Cham. doi.org/10.1007/978-3-031-43458-7_29"@en ;
    foaf:logo <github.com/Naeima/Forest-Observatory-Ontology/
    blob/main/logo.png?raw=true> ;
    dcterms:doi <https://doi.org/10.1007/978-3-031-43458-7_29> ;
    ns2:status "Active" ;
    rdfs:seeAlso <https://naeima.github.io/foo_html/> ;
    rdfs:seeAlso <https://naeima.github.io/foo_html/index.ttl> .

### Provenance Information ###
<https://github.com/Naeima> a prov:Agent,
    foaf:Person ;
    foaf:affiliation "Cardiff University"@en ;
    foaf:mbox <mailto:naeima.hamed@cardiff.ac.uk> ;
    foaf:name "Naeima Hamed"@en .
<https://ontology.forest-observatory.org> a prov:Entity ;
    dc:creator "Data Provider"@en ;
```

```
    dc:description "The dataset from which this ontology was derived."@en ;
    dc:title "Source Dataset"@en ;
    dcterms:created "2023-01-15"^^xsd:date .
<https://link.springer.com/chapter/10.1007/978-3-031-43458-7_29> a prov:Agent,
    foaf:Organization ;
    foaf:homepage <https://rdcu.be/dKNG2> ;
    foaf:name "Springer, Cham"@en .

### FOO Classes (Reused from BBC Wildlife Ontology (wo)) ###
#### BBC Wildlife Ontology Taxonomic Classes ####
wo:Kingdom a owl:Class .
wo:Phylum a owl:Class .
wo:Class a owl:Class .
wo:Order a owl:Class .
wo:Family a owl:Class .
wo:Genus a owl:Class .
wo:Species a owl:Class .
wo:TaxonRank a owl:Class .

#### Kingdom ####
foo:Animalia a owl:Class ;
    rdfs:label "Animalia"@en-gb ;
    rdfs:subClassOf owl:Thing ;
    owl:equivalentClass wo:Animalia ;
    skos:definition "Animalia is the scientific grouping that
    includes all animals.Scientists, historians, and others
    classify similar things together,using a taxonomy."@en.

#### Phylum ####
foo:Chordata a owl:Class ;
    rdfs:label "Chordata"@en-gb ;
    rdfs:subClassOf foo:Animalia ;
    owl:equivalentClass wo:Chordata ;
    skos:definition  "A large phylum of animals that includes
    the vertebrates together with the sea squirts and lancelets.
    They are distinguished by the possession of a notochord
    at some stage during their development."@en .

#### Class ####
foo:Mammalia a owl:Class ;
    rdfs:label "Mammalia"@en-gb ;
    rdfs:subClassOf foo:Chordata ;
    owl:equivalentClass wo:Mammalia ;
    skos:definition  "The highest class of the subphylum
```

```
    Vertebrata comprising humans and all other animals
    that nourish their young with milk secreted by mammary glands."@en.

foo:Reptilia a owl:Class ;
    rdfs:label "Reptilia"@en-gb ;
    rdfs:subClassOf foo:Chordata ;
    owl:equivalentClass wo:Reptilia ;
    skos:definition   "Reptilia is a vertebrate animal
    of a class that includes snakes,
    lizards, crocodiles, turtles, and tortoises.
    They are distinguished by having a dry
    scaly skin and typically laying soft-shelled eggs on land."@en.

#### Order ####
foo:Proboscidea a owl:Class ;
    rdfs:label "Proboscidea"@en-gb ;
    rdfs:subClassOf foo:Mammalia ;
    owl:equivalentClass wo:Proboscidea ;
    skos:definition "Any of an order (Proboscidea)
    of large mammals comprising the elephants and
    extinct related forms."@en.
foo:Carnivora a owl:Class ;
    rdfs:label "Carnivora"@en-gb ;
    rdfs:subClassOf foo:Mammalia ;
    owl:equivalentClass wo:Carnivora ;
    skos:definition  "Carnivora is a monophyletic
    order of placental mammals consisting of the
    most recent common ancestor of all cats and dogs,
    and all descendants of that ancestor.
    Members of this group are formally
    referred to as carnivorans, and have evolved
    to specialize in eating flesh."@en .

foo:Squamata a owl:Class ;
    rdfs:label "Squamata"@en-gb ;
    rdfs:subClassOf foo:Reptilia ;
    owl:equivalentClass wo:Squamata .

#### Family ####
foo:Elephantidae a owl:Class ;
    rdfs:label "Elephantidae"@en-gb ;
    rdfs:subClassOf foo:Proboscidea ;
    owl:equivalentClass wo:Elephantidae ;
    skos:definition "ELEPHANTIDAE is a family of
```

```
    bulky mammals (order Proboscidea) comprising
    the recent elephants and related extinct forms."@en.

#### Additional Families ####

foo:Species  a owl:Class ;
    rdfs:label  "Species"@en-gb ;
    owl:equivalentClass  wo:Species ;
    skos:definition  "Generic class defining a biological species."@en.

foo:Genus  a owl:Class ;
    rdfs:label  "Genus"@en-gb ;
    owl:equivalentClass  wo:Genus .

################################################################################

#### FOO Observation and Sensor Classes (Reused from SOSA) ####

foo:Observation a owl:Class ;
    rdfs:label "Observation"@en ;
    rdfs:definedBy <https://w3id.org/def/foo> ;
    owl:equivalentClass sosa:Observation ;
    skos:definition "Act of carrying out an (Observation) Procedure
    to estimate or calculate a value of a property of a
    FeatureOfInterest (e.g., Elephant)."@en .

foo:Sensor a owl:Class ;
    rdfs:label "Sensor"@en ;
    rdfs:definedBy sosa:Sensor ;
    owl:equivalentClass sosa:Sensor ;
    skos:definition "Device, agent (including humans), or
    software (simulation) involved in, or implementing, a Procedure."@en .

foo:ObservableProperty a owl:Class ;
    rdfs:label "Observable Property"@en ;
    rdfs:definedBy <https://w3id.org/def/foo> ;
    owl:equivalentClass sosa:ObservableProperty ;
    skos:definition "An observable quality (property, characteristic)
    of a FeatureOfInterest."@en .

foo:FeatureOfInterest a owl:Class ;
  rdfs:label "Feature Of Interest"@en ;
  owl:equivalentClass  sosa:FeatureOfInterest ;
  rdfs:isDefinedBy sosa: .
```

```
### FOO Object Properties (Reused from SOSA) ###

foo:hasFeatureOfInterest a owl:ObjectProperty ;
    rdfs:label "has Feature Of Interest"@en ;
    rdfs:comment "A relation between an Observation and the entity
    whose quality was observed."@en ;
    rdfs:domain foo:Observation ;
    rdfs:range foo:FeatureOfInterest ;
    owl:inverseOf sosa:isFeatureOfInterestOf ;
    owl:equivalentProperty sosa:hasFeatureOfInterest .

foo:isFeatureOfInterestOf a owl:ObjectProperty ;
    rdfs:label "is feature of interest of"@en ;
    rdfs:comment "A relation between a FeatureOfInterest
    and an Observation about it."@en ;
    rdfs:domain foo:FeatureOfInterest ;
    rdfs:range foo:Observation ;
    owl:inverseOf sosa:hasFeatureOfInterest ;
    owl:equivalentProperty  sosa:isFeatureOfInterestOf .

foo:madeBySensor a owl:ObjectProperty ;
    rdfs:label "made by sensor"@en ;
    rdfs:comment "Relation between an Observation and the Sensor
    which made the Observation."@en ;
    rdfs:domain foo:Observation ;
    rdfs:range foo:Sensor ;
    owl:inverseOf foo:madeObservation ;
    owl:equivalentProperty sosa:madeBySensor .

foo:observedProperty a owl:ObjectProperty ;
    rdfs:label "observed property"@en ;
    rdfs:comment "Relation linking an Observation to the property
    that was observed."@en ;
    rdfs:domain foo:Observation ;
    rdfs:range foo:ObservableProperty ;
    owl:equivalentProperty    sosa:observedProperty  .

foo:isObservedBy a owl:ObjectProperty ;
    rdfs:label "is observed by"@en ;
    rdfs:comment "Relation between an ObservableProperty and
    the Sensor able to observe it."@en ;
    rdfs:domain foo:ObservableProperty ;
    rdfs:range foo:Sensor ;
```

```
    owl:inverseOf foo:observes ;
    owl:equivalentProperty  sosa:isObservedBy .

foo:madeObservation a owl:ObjectProperty ;
    rdfs:label "made observation"@en ;
    rdfs:comment "Relation between a Sensor and an
    Observation made by the Sensor."@en ;
    rdfs:domain foo:Sensor ;
    rdfs:range foo:Observation ;
    owl:inverseOf sosa:madeBySensor ;
    owl:equivalentProperty  sosa:madeObservation  .

foo:observes a owl:ObjectProperty ;
    rdfs:label "observes"@en ;
    rdfs:comment "Relation between a Sensor and an
    ObservableProperty that it is capable of sensing."@en ;
    rdfs:domain foo:Sensor ;
    rdfs:range foo:ObservableProperty ;
    owl:inverseOf foo:isObservedBy ;
    owl:equivalentProperty  sosa:observes  .

foo:based_near  a owl:ObjectProperty ;
    rdfs:label "near"@en ;
    rdfs:domain foo:FeatureOfInterest ;
    rdfs:range foo:Observation ;
    owl:equivalentProperty    foaf:based_near .

#################################################################
### FOO defined Classes###

foo:Primates a owl:Class ;
    rdfs:label "Primates"@en-gb ;
    rdfs:subClassOf foo:Mammalia ;
    rdfs:definedBy   <http://purl.bioontology.org/ontology/MESH/D011323> ;
    skos:definition  "An order of mammals consisting of more
    than 300 species that include LEMURS; LORISIDAE; TARSIERS; MONKEYS;
    and HOMINIDS. They are characterized by a relatively
    large brain when compared with other terrestrial" .

foo:Cercopithecidae a owl:Class;
    rdfs:label "Cercopithecidae"@en-gb ;
    rdfs:subClassOf foo:Primates ;
    rdfs:definedBy  <purl.bioontology.org/ontology/CSP/0182-1650> .
```

```
foo:Nasalis a  owl:Class ;
    rdfs:label "Nasalis"@en-gb ;
    rdfs:subClassOf foo:Cercopithecidae ;
    skos:definition "Nasalis is a genus within the family Cercopithecidae
    (Old World monkeys), specifically part of the subfamily Colobinae,
    which comprises leaf-eating monkeys.
    The genus Nasalis is characterized by its sole species, Nasalis larvatus,
    commonly known as the proboscis monkey."@en .

foo:NasalisLarvatus a owl:Class ;
    rdfs:label "Proboscis Monkey"@en, "Nasalis larvatus"@la ;
    rdfs:subClassOf foo:Nasalis ;
    owl:equivalentClass  <purl.obolibrary.org/obo/NCBITaxon_43780> .

foo:ElephasMaximus a owl:Class ;
    rdfs:subClassOf foo:Elephantidae, foo:FeatureOfInterest ;
    rdfs:label "Asian Elephant"@en, "Elephas maximus"@la .


foo:Pythonidae a owl:Class ;
    rdfs:subClassOf foo:Squamata   ;
    owl:equivalentClass   <purl.obolibrary.org/obo/NCBITaxon_34984> ;
    rdfs:label "Pythonidae"@en .

foo:Malayopython a owl:Class ;
     rdfs:subClassOf foo:Pythonidae  ;
     owl:equivalentClass  <purl.obolibrary.org/obo/NCBITaxon_1496304> ;
     rdfs:label "Malayopython"@en .

foo:MalayopythonReticulatus a owl:Class ;
    rdfs:subClassOf  foo:Malayopython, foo:FeatureOfInterest  ;
    rdfs:label "Reticulated Python"@en, "Malayopython reticulatus"@la ;
    owl:equivalentClass  <purl.obolibrary.org/obo/NCBITaxon_1496311> ;
    rdfs:definedBy <orca.cardiff.ac.uk/id/eprint/152386/15/2022burgerphd.pdf> .


foo:ManisJavanica a owl:Class ;
    rdfs:subClassOf foo:Mammalia, foo:FeatureOfInterest ;
    rdfs:label "Sunda Pangolin"@en, "Manis javanica"@la ;
    owl:equivalentClass  <purl.bioontology.org/ontology/NCBITAXON/9974> .


### FOO Data Properties ###
```

```
foo:temperature a owl:DatatypeProperty ;
    rdfs:label "Temperature" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:double ;
    skos:definition "Estimated temperature of the elephant in
    Celsius at the moment of data collection." .

foo:count a owl:DatatypeProperty ;
    rdfs:label "Count"@en ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:integer ;
    skos:definition "Observation count per data set." .

foo:cov a owl:DatatypeProperty ;
    rdfs:label "Cov" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:double ;
    skos:definition "TBC" .

foo:direction a owl:DatatypeProperty ;
    rdfs:label "Direction" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:integer ;
    skos:definition "Direction of elephant travel
    at the moment of data collection." .

foo:distance a owl:DatatypeProperty ;
    rdfs:label "Distance"@en ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:double ;
    skos:definition "Distance (m) travelled from the last to the
    current data collection point." .

foo:gMTDate a owl:DatatypeProperty ;
    rdfs:label "GMT Date" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:date ;
    skos:definition "The GMT date in Sabah, Malaysia, when the
    GPS collar records its readings." .

foo:gMTTime a owl:DatatypeProperty ;
    rdfs:label "GMT Time" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:time ;
```

```
    skos:definition "The GMT time in Sabah, Malaysia, when the
    GPS collar records its readings." .

foo:hDOP a owl:DatatypeProperty ;
    rdfs:label "HDOP" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:double ;
    skos:definition "Horizontal Dilution of Precision (HDOP),
    indicating GPS accuracy." .

foo:horizon a owl:DatatypeProperty ;
    rdfs:label "Horizon"@en ;
    rdfs:domain foo:soilObservation ;
    rdfs:range xsd:string ;
    skos:definition "Soil horizon sampled."@en .

foo:id a owl:DatatypeProperty ;
    rdfs:label "id"@en ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:string .

foo:landUse a owl:DatatypeProperty ;
    rdfs:label "Land Use"@en ;
    rdfs:domain foo:soilObservation ;
    rdfs:range xsd:string ;
    skos:definition "Land use of the study plots."@en .

foo:latitude a owl:DatatypeProperty ;
    rdfs:label "Latitude" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:double ;
    owl:equivalentProperty pos:lat ;
    skos:definition "Latitudinal coordinate of the elephant." .

foo:localDate a owl:DatatypeProperty ;
    rdfs:label "Local Date" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:date ;
    skos:definition "The local date in Sabah, Malaysia." .

foo:localTime a owl:DatatypeProperty ;
    rdfs:label "Local Time" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:time ;
```

```
    skos:definition "The local time in Sabah, Malaysia." .

foo:longitude a owl:DatatypeProperty ;
    rdfs:label "Longitude" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:double ;
    owl:equivalentProperty pos:long ;
    skos:definition "Longitudinal coordinate of the elephant." .

foo:speed a owl:DatatypeProperty ;
    rdfs:label "Speed" ;
    rdfs:domain foo:gPSObservation ;
    rdfs:range xsd:double ;
    skos:definition "Speed of the elephant at the moment of data collection." .

### Soil Data Properties ###

foo:clay a owl:DatatypeProperty ;
    rdfs:label "Clay"@en ;
    rdfs:domain foo:soilObservation  ;
    rdfs:range xsd:double ;
    skos:definition "Clay content of the soil sample."@en .

foo:silt a owl:DatatypeProperty ;
    rdfs:label "Silt"@en ;
    rdfs:domain foo:soilObservation ;
    rdfs:range xsd:double ;
    skos:definition "Silt content of the soil sample."@en .

foo:site a owl:DatatypeProperty ;
    rdfs:label "Site"@en ;
    rdfs:domain foo:soilObservation ;
    rdfs:range xsd:string ;
    skos:definition "Geographical area/site which samples were taken from."@en .

foo:soilPH a owl:DatatypeProperty ;
    rdfs:label "Soil PH" ;
    rdfs:domain foo:soilObservation ;
    rdfs:range xsd:double ;
    skos:definition "Measured pH of the soil sample."@en .

foo:subplot a owl:DatatypeProperty ;
    rdfs:label "subPlot"@en ;
    rdfs:domain foo:gPSObservation ;
```

```
    rdfs:range xsd:string ;
    skos:definition "Number of subplot sampled within each 1 Ha plot."@en .

foo:totalC a owl:DatatypeProperty ;
    rdfs:label "Total C"@en ;
    rdfs:domain foo:soilObservation ;
    rdfs:range xsd:double ;
    skos:definition "Total carbon content of the soil sample."@en .

foo:totalN a owl:DatatypeProperty ;
    rdfs:label "Total N"@en ;
    rdfs:domain foo:soilObservation ;
    rdfs:range xsd:double ;
    skos:definition "Total nitrogen content of the soil sample."@en .

### Tree Observation Data Properties ###

foo:lianaDBH_cm a owl:DatatypeProperty ;
    rdfs:label "lianaDBH_cm 10a"@en ;
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:string .

 foo:subplotRadius_m a owl:DatatypeProperty ;
    rdfs:label "SubplotRadius_m 30"@en ;
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:float .

 foo:treeDBH_cm a owl:DatatypeProperty ;
    rdfs:label "TreeDBH_cm 110"@en ;
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:float .

 foo:treeHeight_m   a owl:DatatypeProperty ;
    rdfs:label "treeHeight_m 60"@en ;
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:float .

 foo:treeID  a owl:DatatypeProperty ;
    rdfs:label "TreeID"@en ;
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:string .

 foo:treeDBH_cm  a owl:DatatypeProperty ;
    rdfs:label "TreeDBH_cm 110"@en ;
```

```
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:float .

 foo:treeIndividualNo  a owl:DatatypeProperty ;
    rdfs:label "TreeIndividualNo"@en ;
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:integer .

 foo:treeIndividualNo  a owl:DatatypeProperty ;
    rdfs:label "TreeIndividualNo"@en ;
    rdfs:domain foo:treeObservation ;
    rdfs:range xsd:integer .


### Camera Trap Image Data Properties ###
foo:name a owl:DatatypeProperty ;
    rdfs:label "Image Name"@en ;
    rdfs:domain foo:imageObservation ;
    rdfs:range xsd:string ;
    skos:definition "Name assigned to an image at collection time."@en .

foo:path a owl:DatatypeProperty ;
    rdfs:label "Image Path"@en ;
    rdfs:domain foo:imageObservation ;
    rdfs:range xsd:anyURI ;
    skos:definition "The URI pointing to the location of
    the image in secure cloud storage."@en .

foo:localDate a owl:DatatypeProperty ;
    rdfs:label "Local Date"@en ;
    rdfs:domain foo:imageObservation;
    rdfs:range xsd:date ;
    skos:definition "Current local date in Sabah,
    Malaysia when the GPS collar collects its readings."@en .

foo:localTime a owl:DatatypeProperty ;
    rdfs:label "Local Time"@en ;
    rdfs:domain foo:imageObservation ;
    rdfs:range xsd:time ;
    skos:definition "The current local time in Sabah,
    Malaysia when the GPS collar collects its readings."@en .

foo:gMTDate a owl:DatatypeProperty ;
    rdfs:label "GMT Date"@en ;
```

```
    rdfs:domain foo:imageObservation ;
    rdfs:range xsd:date ;
    skos:definition "The GMT date in Sabah, Malaysia
    when the GPS collar collects its readings."@en .

foo:gMTTime a owl:DatatypeProperty ;
    rdfs:label "GMT Time"@en ;
    rdfs:domain foo:imageObservation ;
    rdfs:range xsd:time ;
    skos:definition "The GMT time in Sabah, Malaysia
    when the GPS collar collects its readings."@en .

foo:model a owl:DatatypeProperty ;
    rdfs:label "Camera Model"@en ;
    rdfs:domain foo:imageObservation ;
    rdfs:range xsd:string ;
    skos:definition "The model of the trail camera
    used to capture the image."@en .

foo:make a owl:DatatypeProperty ;
    rdfs:label "Camera Make"@en ;
    rdfs:domain foo:imageObservation ;
    rdfs:range xsd:string ;
    skos:definition "The make of the trail camera
    used to capture the image."@en .

foo:imageFile a owl:DatatypeProperty ;
    rdfs:label "Image File"@en ;
    rdfs:domain  foo:imageObservation ;
    rdfs:range    xsd:string ;
    skos:definition "The image file name generated by
    the image observation."@en .

foo:cameraLocation a owl:DatatypeProperty ;
    rdfs:label "Camera Location"@en ;
    rdfs:domain  foo:imageObservation ;
    rdfs:range    xsd:string ;
    skos:definition "The location information
    (address) of the camera trap."@en .

 foo:animalDetected a owl:DatatypeProperty ;
    rdfs:label "Animal Detected"@en ;
    rdfs:domain  foo:imageObservation ;
    rdfs:range    xsd:string .
```

```
### FOO Instances ###
#### Sensor Instances ####
foo:aqeelaGPS a owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Aqeela GPS"@en ;
    foo:hasFeatureOfInterest foo:Aqeela ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Aqeela."@en ;
    foo:observes foo:gPSObservation .

foo:bikang1GPS a owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Bikang 1 GPS"@en ;
    foo:hasFeatureOfInterest foo:Bikang1 ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Bikang 1."@en ;
    foo:observes foo:gPSObservation .

foo:bikang2GPS a owl:NamedIndividual,  foo:Sensor ;
    rdfs:label "Bikang 2 GPS"@en ;
    foo:hasFeatureOfInterest foo:Bikang2 ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Bikang 2."@en ;
    foo:observes foo:gPSObservation .

foo:binbinganGPS a owl:NamedIndividual,  foo:Sensor ;
    rdfs:label "Binbingan GPS"@en ;
    foo:hasFeatureOfInterest foo:Binbingan ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Binbingan."@en ;
    foo:observes foo:gPSObservation .

foo:guliGPS a  owl:NamedIndividual,  foo:Sensor ;
    rdfs:label "Guli GPS"@en ;
    foo:hasFeatureOfInterest foo:Guli ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Guli."@en ;
    foo:observes foo:gPSObservation .

foo:itaGPS a owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Ita GPS"@en ;
    foo:hasFeatureOfInterest foo:Ita;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Ita."@en ;
    foo:observes foo:gPSObservation .
```

```
foo:jasminGPS a owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Jasmin GPS"@en ;
    foo:hasFeatureOfInterest foo:Jasmin ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Jasmin."@en ;
    foo:observes foo:gPSObservation .

foo:jasperGPS a owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Jasper GPS"@en ;
    foo:hasFeatureOfInterest foo:Jasper ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Jasper."@en ;
    foo:observes foo:gPSObservation .

foo:kasihGPS a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Kasih GPS"@en ;
    foo:hasFeatureOfInterest foo:Kasih ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Kasih."@en ;
    foo:observes foo:gPSObservation .

foo:kumaGPS a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Kuma GPS"@en ;
    foo:hasFeatureOfInterest foo:Kuma ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Kuma."@en ;
    foo:observes foo:gPSObservation .

foo:liunGPS a  owl:NamedIndividual,  foo:Sensor ;
    rdfs:label "Liun GPS"@en ;
    foo:hasFeatureOfInterest foo:Luin ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Liun."@en ;
    foo:observes foo:gPSObservation .

foo:maliauGPS a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Maliau GPS"@en ;
    foo:hasFeatureOfInterest foo:Maliau ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Maliau."@en ;
    foo:observes foo:gPSObservation .

foo:merotaiGPS a  owl:NamedIndividual, foo:Sensor ;
```

```
    rdfs:label "Merotai GPS"@en ;
    foo:hasFeatureOfInterest foo:Merotai ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Merotai."@en ;
    foo:observes foo:gPSObservation .

foo:puteriGPS a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Puteri GPS"@en ;
    foo:hasFeatureOfInterest foo:Puteri ;
     skos:definition "A GPS collar sensor fitted around the neck of an
     Asian elephant named Puteri."@en ;
    foo:observes foo:gPSObservation .

foo:pututGPS a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Putut GPS"@en ;
    foo:hasFeatureOfInterest foo:Putut ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Putut."@en ;
    foo:observes foo:gPSObservation .

foo:sejatiGPS a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Sejati GPS"@en ;
    foo:hasFeatureOfInterest foo:Sejati ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Sejati."@en ;
    foo:observes foo:gPSObservation .

foo:seriGPS a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Seri GPS"@en ;
    foo:hasFeatureOfInterest foo:Seri ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Seri."@en ;
    foo:observes foo:gPSObservation .

foo:tulidGPS a  owl:NamedIndividual,  foo:Sensor ;
    rdfs:label "Tulid GPS"@en ;
    foo:hasFeatureOfInterest foo:Tulid ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Tulid."@en ;
    foo:observes foo:gPSObservation .

foo:tunglapGPS a   owl:NamedIndividual,  foo:Sensor ;
    rdfs:label "Tunglap GPS"@en ;
    foo:hasFeatureOfInterest foo:Tunglap ;
```

```
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Tunglap."@en ;
    foo:observes foo:gPSObservation .

foo:umas2GPS a   owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Umas2 GPS"@en ;
    foo:hasFeatureOfInterest foo:Umas2 ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Umas2."@en ;
    foo:observes foo:gPSObservation .

foo:daraGPS a   owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Dara GPS"@en ;
    foo:hasFeatureOfInterest foo:Dara ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Dara."@en ;
    foo:observes foo:gPSObservation .

foo:abawGPS a   owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Abaw GPS"@en ;
    foo:hasFeatureOfInterest foo:Abaw ;
    skos:definition "A GPS collar sensor fitted around the neck of an
    Asian elephant named Abaw ."@en ;
    foo:observes foo:gPSObservation .


#### Animal Instances ####
foo:aqeela a owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Aqeela"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:guli a   owl:NamedIndividual, foo:ElephasMaximus;
    rdfs:label "Guli"@en ;
    skos:definition "Male Asian Elephant."@en .

foo:bikang1 a   owl:NamedIndividual, foo:ElephasMaximus;
    rdfs:label "Bikang 1"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:bikang2 a   owl:NamedIndividual, foo:ElephasMaximus;
    rdfs:label "Bikang 2"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:dara a   owl:NamedIndividual, foo:ElephasMaximus;
```

```
    rdfs:label "Dara"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:abaw a   owl:NamedIndividual, foo:ElephasMaximus;
    rdfs:label "Abaw"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:ita a   owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Ita"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:jasmin a  owl:NamedIndividual,  foo:ElephasMaximus ;
    rdfs:label "Jasmin"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:jasper a  owl:NamedIndividual,  foo:ElephasMaximus ;
    rdfs:label "Jasper"@en ;
    skos:definition "Male Asian Elephant."@en .

foo:kasih a owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Kasih"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:kuma a  owl:NamedIndividual,  foo:ElephasMaximus ;
    rdfs:label "Kuma"@en ;
    skos:definition "Male Asian Elephant."@en .

foo:liun a  owl:NamedIndividual,  foo:ElephasMaximus ;
    rdfs:label "Liun"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:maliau a  owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Maliau"@en ;
    skos:definition "Male Asian Elephant."@en .

foo:merotai a   owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Merotai"@en ;
    skos:definition "Male Asian Elephant."@en .

foo:puteri a  owl:NamedIndividual, foo:ElephasMaximus;
    rdfs:label "Puteri"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:putut a  owl:NamedIndividual, foo:ElephasMaximus ;
```

```
    rdfs:label "Putut"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:sejati a   owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Sejati"@en ;
    skos:definition "Male Asian Elephant."@en .

foo:seri a   owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Seri"@en ;
    skos:definition "Female Asian Elephant ."@en .

foo:tulid a   owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Tulid"@en ;
    skos:definition "Female Asian Elephant ."@en .

foo:tunglap a  owl:NamedIndividual,  foo:ElephasMaximus ;
    rdfs:label "Tunglap"@en ;
    skos:definition "Female Asian Elephant."@en .

foo:umas2 a   owl:NamedIndividual, foo:ElephasMaximus ;
    rdfs:label "Umas2"@en ;
    skos:definition "Male Asian Elephant ."@en .

foo:gPSObservation a   owl:NamedIndividual,  foo:Observation ;
    rdfs:label "GPS Observation"@en ;
    foo:observedProperty  foo:id, foo:altitude , foo:count , foo:cov ,
    foo:direction , foo:distance , foo:gMTDate ,
    foo:gMTTime , foo:hDOP , foo:latitude , foo:localDate ,
    foo:local-time  , foo:longitude , foo:speed , foo:temperature  ;
    foo:hasFeatureOfInterest foo:ElephasMaximus ;
    foo:madeBySensor foo:AqeelaGPS, foo:Bikang1GPS,
    foo:Bikang2GPS, foo:BinbinganGPS, foo:DaraGPS, foo:GuliGPS,
    foo:ItaGPS, foo:JasminGPS, foo:JasperGPS, foo:KasihGPS,
    foo:KumaGPS, foo:LiunGPS, foo:MaliauGPS, foo:MerotaiGPS,
    foo:PuteriGPS, foo:PututGPS, foo:SejatiGPS, foo:SeriGPS,
    foo:TulidGPS, foo:TunglapGPS, foo:Umas2GPS ;
    foo:resultTime "2011-10-26T07:40:35"^^xsd:dateTime,
    "2015-10-26T07:40:35"^^xsd:dateTime .

####Soil Modeling ####
foo:Soil a owl:Class ;
    rdfs:label "Soil"@en ;
    rdfs:subClassOf   foo:FeatureOfInterest ;
    owl:equivalentClass  <saref.etsi.org/saref4agri/Soil>  .
```

```
#### Soil Sensor ####
foo:soilSensor a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Soil Sensor"@en ;
    owl:sameAs  <saref.etsi.org/saref4agri/SoilTensiometer> ;
    foo:hasFeatureOfInterest foo:Soil .

#### Soil Observation ####
foo:soilObservation a  owl:NamedIndividual,  foo:Observation ;
    rdfs:label "Soil Observation"@en ;
# Soil properties observed
    foo:observedProperty foo:cNRatio ,
    foo:clay,
    foo:horizon ,
    foo:identifier ,
    foo:inorganicP ,
    foo:landUse ,
    foo:plotName ,
    foo:sand ,
    foo:silt  ,
    foo:site  ,
    foo:soilPH  ,
    foo:subplot  ,
    foo:totalC  ,
    foo:totalN ,
    foo:totalP ;
# Link the observation to the soil feature and sensor
    foo:hasFeatureOfInterest foo:Soil ;
    foo:madeBySensor foo:soilSensor .
#### Tree Modeling ####
foo:Tree a owl:Class ;
    rdfs:subClassOf   foo:FeatureOfInterest ;
    owl:equivalentClass  <purl.dataone.org/odo/ECSO_00000501> ;
    rdfs:label "Tree"@en .

#### Tree Observation ####
foo:treeSensor a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Tree Sensor"@en ;
    foo:observes foo:treeProperties ;
    foo:hasFeatureOfInterest foo:Tree .

foo:treeObservation a  owl:NamedIndividual, foo:Observation ;
    rdfs:label "Tree Observation"@en ;
```

```
# Observation metadata
    foo:observedProperty  foo:date ,
    foo:iD ,
    # Tree measurements
    foo:lianaDBH_cm ,
    foo:subplotRadius_m ,
    foo:treeDBH_cm  ,
    foo:treeHeight_m  ,
    foo:treeID  ,
    foo:treeIndividualNo  ,
    foo:treeNLianas ,
    foo:treeNotes ;

# Link the observation to the feature of interest (tree) and sensor
    foo:hasFeatureOfInterest foo:Tree ;
    foo:madeBySensor foo:treeSensor .

#### Lianas as Feature of Interest ####
foo:lianas a  owl:NamedIndividual, foo:Tree ;
    rdfs:label "Lianas"@en ;
    foo:isObservedBy foo:lianaSensor .

#### Liana Sensor ####
foo:lianaSensor a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Liana Sensor"@en ;
    foo:observes foo:lianaProperties .

#### Grow Borneo Project ####
foo:Project a owl:Class ;
    rdfs:label "Project" ;
    rdfs:comment "Represents a reforestation project." .

### Object Properties
foo:isPlantedIn a owl:ObjectProperty ;
    rdfs:label "is planted in" ;
    rdfs:domain foo:Tree ;
    rdfs:range foo:Project ;
    rdfs:comment "Links a tree species to the reforestation project
    where it's planted." .

### Individual (Grow Borneo Project)
foo:growBorneo a owl:NamedIndividual, foo:Project ;
    rdfs:label "Grow Borneo" ;
    rdfs:comment "A reforestation project in Borneo planting
```

```
    various tree species." .

### Link Tree Species to Grow Borneo
foo:bongkol a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Bongkol" ;
    rdfs:comment "A tree species named Bongkol in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:selongapid a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Selongapid" ;
    rdfs:comment "A tree species named Selongapid in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:binuang a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Binuang" ;
    rdfs:comment "A tree species named Binuang in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:terosob a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Terosob" ;
    rdfs:comment "A tree species named Terosob in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:kelumpang a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Kelumpang" ;
    rdfs:comment "A tree species named Kelumpang in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:mangkapon a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Mangkapon" ;
    rdfs:comment "A tree species named Mangkapon in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:nyatoh a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Nyatoh" ;
    rdfs:comment "A tree species named Nyatoh in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:durian a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Durian" ;
    rdfs:comment "A tree species named Durian in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:tarap a owl:NamedIndividual, foo:Tree ;
```

```
    rdfs:label "Tarap" ;
    rdfs:comment "A tree species named Tarap in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:rambutan a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Rambutan" ;
    rdfs:comment "A tree species named Rambutan in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:pulai a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Pulai" ;
    rdfs:comment "A tree species named Pulai in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:payungPayung a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Payung Payung" ;
    rdfs:comment "A tree species named Payung Payung in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:kayuMalam a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Kayu Malam" ;
    rdfs:comment "A tree species named Kayu Malam in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:kerodong a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Kerodong" ;
    rdfs:comment "A tree species named Kerodong in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:keruingPaya a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Keruing Paya" ;
    rdfs:comment "A tree species named Keruing Paya in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:bayur a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Bayur" ;
    rdfs:comment "A tree species named Bayur in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:tangkol a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Tangkol" ;
    rdfs:comment "A tree species named Tangkol in Malay." ;
    foo:isPlantedIn foo:growBorneo .
```

```
foo:sepat a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Sepat" ;
    rdfs:comment "A tree species named Sepat in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:belian a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Belian" ;
    rdfs:comment "A tree species named Belian in Malay." ;
    foo:isPlantedIn foo:growBorneo .

foo:keranji a owl:NamedIndividual, foo:Tree ;
    rdfs:label "Keranji" ;
    rdfs:comment "A tree species named Keranji in Malay." ;
    foo:isPlantedIn foo:growBorneo .
### Individual (Grow Borneo Project)
foo:growBorneo a  owl:NamedIndividual,  foo:Project ;
    rdfs:label "Grow Borneo" ;
    rdfs:comment "A reforestation project in Borneo that
    plants various tree species." .

##### Camera Trap Images Modeling ####
#### Image as a Feature of Interest ####
foo:Image a owl:Class ;
    rdfs:subClassOf   foo:FeatureOfInterest ;
    owl:equivalentClass   foaf:Image ;
    rdfs:label "Camera Trap Image"@en ;
    skos:definition "Image generated by motion-activated
    wildlife cameras."@en ;
    rdfs:comment   "The image as feature of interest for the
    camera trap because it carries data critical to wildlife
    analysis such as species.
    However, it should ideally have a clear, semantically relevant
    role--such as representing visual evidence in an image
    recognition or object detection model--otherwise,
    it might dilute the clarity of the knowledge graph or ontology.";
    rdfs:definedBy <http://w3id.org/def/foo#> .

#### Camera Trap Sensor ####
foo:cameraTrap a  owl:NamedIndividual, foo:Sensor ;
    rdfs:label "Camera Trap"@en ;
    foo:observes foo:imageObservation ;
    foo:observedProperty  foo:model  ,
    foo:make  ;
    foo:hasFeatureOfInterest foo:Image .
```

```
#### Image Observation ####
foo:imageObservation a owl:NamedIndividual,  foo:Observation ;
    rdfs:label "Image Observation"@en ;
    foo:hasFeatureOfInterest foo:Image ;
    foo:madeBySensor foo:cameraTrap ;
    foo:observedProperty   foo:imageFile ,
    foo:cameraLocation ,
    foo:animalDetected .

##### Oil Palm Plantation ######
foo:OilPalmPlantation a  owl:Class ;
    rdfs:label   "Oil Palm Plantation" ;
    rdfs:subClassOf  foo:FeatureOfInterest ;
    rdfs:comment  "Oil palm plantations near the
    Danau Girang Field Centre (DGFC) in Sabah, Malaysia,
    are situated within the fragmented landscape of the
    Lower Kinabatangan floodplain, approximately between
    5.4°N to 5.6°N latitude and 117.9°E to 118.1°E longitude.
    This region includes a mix of protected forests,
    degraded habitats, and extensive plantations,
    #often bordering riparian corridors along the Kinabatangan River.
    These plantations have significantly impacted biodiversity
    and habitat connectivity, posing challenges for
    wildlife such as Bornean elephants and orangutans.";
    owl:equivalentClass   <purl.obolibrary.org/obo/ENVO_00000120> .

foo:plantation a owl:NamedIndividual, foo:OilPalmPlantation;
    pos:latitude "5.36"^^xsd:float;
    pos:longitude "118.66 "^^xsd:float.

#### University and Danau Girang Field Centre ####

# Define University as a subclass of foaf:Organization
foo:University rdf:type owl:Class ;
    rdfs:label "University"@en ;
    rdfs:comment "A subclass of FOAF's Organization
    representing academic institutions involved in
    research and higher education."@en ;
    rdfs:subClassOf foaf:Organization ;
    dcterms:description "Universities grant academic
    degrees and conduct research, often partnering
    with other organizations for projects like
    wildlife conservation."@en .
```

```
foo:WildlifeDepartment rdf:type owl:Class ;
    rdfs:label "Wildlife Department"@en ;
    rdfs:comment "A government or non-government
    organization responsible for managing and
    conserving wildlife and their habitats."@en ;
    dcterms:description "Wildlife Departments
    oversee policies, conservation programs,
    and research to protect wildlife and their ecosystems."@en .
foo:FieldCentre rdf:type owl:Class ;
    rdfs:label "Field Centre"@en ;
    rdfs:comment "A facility dedicated to supporting research,
    conservation, and education in specific ecological
    or wildlife domains."@en ;
    dcterms:description "Field Centres provide infrastructure
    and expertise for field research and conservation activities,
    often in partnership with other organizations."@en .
# Properties of a Field Centre and other buildings
foo:location rdf:type owl:ObjectProperty ;
    rdfs:domain foo:FieldCentre ;
    rdfs:range rdfs:Literal ;
    rdfs:label "Location"@en ;
    rdfs:comment "Specifies the geographical location of an entity."@en .

foo:supportedBy rdf:type owl:ObjectProperty ;
    rdfs:domain foo:FieldCentre ;
    rdfs:range foo:University ;
    rdfs:label "Supported By"@en ;
    rdfs:comment "Organizations or entities providing
    financial, technical, or logistical support for the Field Centre."@en .

foo:focusArea rdf:type owl:DatatypeProperty ;
    rdfs:domain foo:FieldCentre ;
    rdfs:range rdfs:Literal ;
    rdfs:label "Focus Area"@en ;
    rdfs:comment "The main area of research or
    conservation focus for the Field Centre."@en .

foo:mission rdf:type owl:DatatypeProperty ;
    rdfs:domain foo:FieldCentre ;
    rdfs:range rdfs:Literal ;
    rdfs:label "Mission Statement"@en ;
    rdfs:comment "The overarching goal or
    purpose of the Field Centre."@en .
# Supporting Institutions
```

```
foo:cardiffUniversity rdf:type foo:University ;
    rdfs:label "Cardiff University"@en ;
    foo:location "Wales, UK" .

foo:sabahWildlifeDepartment rdf:type foo:WildlifeDepartment ;
    rdfs:label "Sabah Wildlife Department"@en ;
    foo:location "Borneo, Malaysia" .

# DGFC
foo:danauGirangFieldCentre rdf:type foo:FieldCentre ;
    rdfs:label "Danau Girang Field Centre"@en ;
    dcterms:description "A field centre focused on wildlife research
    and conservation in the Lower Kinabatangan Wildlife Sanctuary,
    supported by Cardiff University and the Sabah Wildlife Department."@en ;
    foo:location "Lower Kinabatangan Wildlife Sanctuary, Sabah, Malaysia" ;
    foo:supportedBy foo:cardiffUniversity, foo:sabahWildlifeDepartment ;
    foo:focusArea "Conservation Research, Wildlife Studies,
    Fragmented Landscapes"@en ;
    foo:mission "Support Sabah's conservation priorities and
    enhance understanding of wildlife issues in fragmented
    landscapes through research."@en .
##### How to reuse FOO #######
# To enable effective reuse of this ontology, please follow these guidelines:
# 1. Create your own custom ontology using identical class and property
names as in this model.
# 2. Import the `FOO` ontology (https://w3id.org/def/foo#) directly into
the new ontology.
# 3. For each class and property, link your custom definitions to
those in FOO.
# using owl:equivalentClass, owl:equivalentProperty, and owl:sameAs.
# This approach will maintain semantic consistency and allow for
smooth interoperability
# across ontologies that reference shared terms and structures.
```

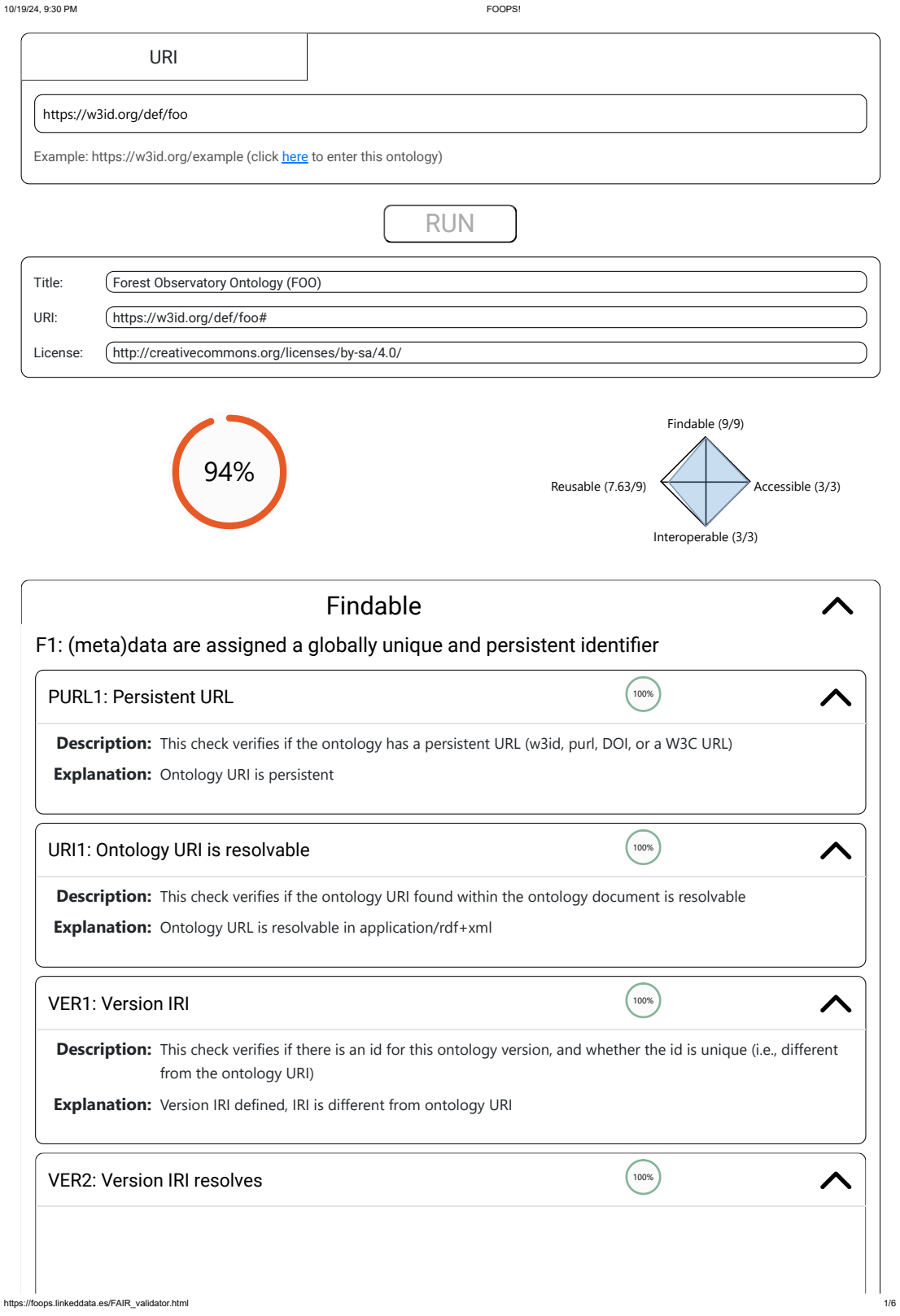


Figure 1 FOOPS! Score

Description: This check verifies if the version IRI resolves

Explanation: Version IRI resolves

URI2: Consistent ontology IDs 100%

Description: This check verifies if the ontology URI is equal to the ontology ID

Explanation: Ontology URI is equal to ontology id

F2: data are described with rich metadata (defined by R1 below)

OM1: Minimum metadata 100%

Description: This check verifies if the The following minimum metadata [title, description, license, version iri, creator, creationDate, namespace URI] are present in the ontology

Explanation: All the minimum metadata were found!

F3: metadata clearly and explicitly include the identifier of the data it describes

FIND1: Ontology prefix 100%

Description: This check verifies if an ontology prefix is available

Explanation: Prefix declaration found in the ontology: foo

F4: (meta)data are registered or indexed in a searchable resource

FIND2: Prefix is in registry 100%

Description: This check verifies if the ontology prefix can be found in prefix.cc or LOV registries. This check also verifies if the prefix resolves to the same namespaceprefix found in the ontology.

Explanation: Prefix declaration found with correct namespace (in LOV)

FIND3: Ontology in metadata registry 100%

Description: This check verifies if the ontology can be found in a public registry (LOV)

Explanation: Ontology namespace found in LOV repository

Accessible

A1: (meta)data are retrievable by their identifier using a standardized communications protocol

CN1: Content negotiation for RDF and HTML 100%

Description: This check verifies of the ontology URI is published following the right content negotiation for RDF and HTML

Explanation: Ontology available in: HTML, RDF

A2: metadata are accessible, even when the data are no longer available

FIND_3_BIS: Metadata are accessible, even when ontology is not 100%

Description: Metadata are accessible even when the ontology is no longer available. Since the metadata is usually included in the ontology, this check verifies whether the ontology is registered in a public metadata registry (LOV)

Explanation: Ontology namespace found in LOV repository

A1.1: the protocol is open, free, and universally implementable

HTTP1: Open protocol 100%

Description: This check verifies if the ontology uses an open protocol (HTTP or HTTPS)

Explanation: The ontology uses an open protocol

Interoperable

I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

RDF1: RDF Availability 100%

Description: This check verifies if the ontology has an RDF serialization (ttl, n3, rdf/xml, json-ld)

Explanation: Ontology available in RDF

I2: (meta)data use vocabularies that follow FAIR principles

VOC1: Vocabulary reuse (metadata) 100%

Description: This check verifies if the ontology reuses other vocabularies for declaring metadata terms

Explanation: Ontology reuses existing vocabularies for declaring metadata.

Imported/Reused URIs:

- http://purl.org/dc/elements/1.1/
- http://purl.org/dc/terms/
- http://purl.org/vocab/vann/
- http://www.w3.org/2000/01/rdf-schema#
- http://www.w3.org/2002/07/owl#
- http://www.w3.org/ns/prov#
- http://xmlns.com/foaf/0.1/

- https://schema.org/

VOC2: Vocabulary reuse 100%

Description: This check verifies if the ontology imports/extends other vocabularies (besides RDF, OWL and RDFS)

Explanation: The ontology imports the following vocabularies:

Imported/Reused URIs:

- http://www.w3.org/ns/sosa/

Reusable

R1: meta(data) are richly described with a plurality of accurate and relevant attributes

DOC1: HTML availability 100%

Description: This check verifies if the ontology has an HTML documentation

Explanation: Ontology available in HTML

OM2: Recommended metadata 100%

Description: This check verifies if the following recommended metadata [NS Prefix, version info, creation date, citation] are present in the ontology. It also checks if [contributor] is present, but with no penalty (as no all ontologies may have a contributor)

Explanation: All recommended metadata found!. Warning: The following OPTIONAL recommended metadata could not be found: contributor. Please consider adding them if appropriate.

OM3: Detailed metadata 50%

Description: This check verifies if the following detailed metadata [doi, publisher, logo, status, source, issued date] are present in the ontology. It also checks if [previous version, backward compatibility, modified] are present, but with no penalty (as no all ontologies may have, e.g., a previous version)

Explanation: The following metadata was not found: doi, status, source. Warning: The following OPTIONAL detailed metadata could not be found: previous version, backwards compatibility, modified. Please consider adding them if appropriate.

VOC3: Documentation: labels 97%

Description: This check verifies the extent to which all ontology terms have labels (rdfs:label in OWL vocabularies, skos:prefLabel in SKOS vocabularies)

Explanation: Labels found for 68 out of 70 terms.

Affected URIs:

- https://w3id.org/def/foo#Family
- https://w3id.org/def/foo#Mammilia

VOC4: Documentation: definitions 16%

Description: This check verifies whether all ontology terms have descriptions (rdfs:comment in OWL vocabularies, skos:definition in SKOS vocabularies)

Explanation: Descriptions found for 11 out of 70 terms

Affected URIs:

- https://w3id.org/def/foo#Animalia
- https://w3id.org/def/foo#Carnivora
- https://w3id.org/def/foo#Cercopithecidae
- https://w3id.org/def/foo#Chordata
- https://w3id.org/def/foo#Elephantidae

Show more

R1.1: (meta)data are released with a clear and accessible data usage license

OM4.1: License availability 100%

Description: This check verifies if a license associated with the ontology

Explanation: A license was found http://creativecommons.org/licenses/by-sa/4.0/

OM4.2: License is resolvable 100%

Description: This check verifies if the ontology license is resolvable

Explanation: License could be resolved

R1.2: (meta)data are associated with detailed provenance

OM5_1: Basic provenance metadata 100%

Description: This check verifies if basic provenance is available for the ontology: [author, creation date]. This check also verifies whether [contributor, previous version] are present, but with no penalty (as no all ontologies may have a previous version or a contributor)

Explanation: All basic provenance metadata found!. Warning: The following OPTIONAL provenance metadata could not be found: contributor, previous version. Please consider adding them if appropriate.

OM5_2: Detailed provenance metadata 100%

Description: This check verifies if detailed provenance information is available for the ontology: [issued date, publisher]

Explanation: All detailed provenance metadata found!

Daniel Garijo & María Poveda-Villalón
Contact email: foops@delicias.dia.fi.upm.es
Built with Bootstrap
Latest revision July, 2021