FOODS: Ontology-based Knowledge Graphs for Forest Observatories

Wildlife research activities generate data on ecosystems and species interactions from varied independent projects. Forest Observatories are online platforms that curate, integrate, and analyze wildlife research data for forest monitoring. However, integrating data from disparate sources can be challenging due to data heterogeneity. This study, in collaboration with a research facility in the forest of Sabah, Malaysian Borneo, proposes a novel approach to integrate heterogeneous wildlife data for Forest Observatories. We used the Forest Observatory Ontology (FOO) to standardize wildlife data entities generated by sensors. Four semantically modeled wildlife datasets populated FOO, resulting in an ontology-based knowledge graph named FOODS (Forest Observatory Ontology Data Store). We evaluated FOO and FOODS using specialized open-source ontology scanners, domain experts' feedback, and applied use cases. This study contributes FOODS, the first ontology-based knowledge graph for Forest Observatories, which provides accurate query responses, reasoning about data, and granular data acquisition from diverse datasets. FOO documentation and FOODS resources are available at https://w3id.org/xxx/xxx* and https://xxxxxxxx.xxxxxx-xxxxxxxxxxx.org.

CCS Concepts: • **Information systems** → **Graph-based database models**; **Information retrieval query processing**.

Additional Key Words and Phrases: Wildlife data, Internet of Things, Ontology, Knowledge Graph

ACM Reference Format:
. 2024. FOODS: Ontology-based Knowledge Graphs for Forest Observatories. 1, 1 (July 2024), 41 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Forest Observatories integrate and analyze wildlife data to answer questions that support data-driven analysis and forest monitoring [53]. Such observatories can enhance the understanding of ecosystems, species interactions, and environmental changes, aiding conservation efforts and informed decision-making [29]. 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 [92].

Siloed data hinder collaboration as groups work independently, thereby reducing opportunities for data sharing [71]. For example, consider one group studying the impact of elephant populations on soil health in a specific ecosystem, whereas another group investigating the behavior 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

Author's address:

XXXX-XXXX/2024/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 behavior. Collaboration between these two groups can be facilitated by using a common data store that standardizes the datasets and links their entities. This online 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 [76, 105, 114].

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 exist in past studies [18, 19, 56]. However, these approaches often lack meaningful connections between the data entities. Data scientists can derive substantial benefits from incorporating Semantic Web technologies 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 [49] 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 [52], 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 foundational 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. Our study's contributions to Forest Observatories include the following:

(1) The Forest Observatory Ontology (FOO) and its knowledge graphs, equipped with online documentation for describing wildlife data generated by sensors.
(2) A resource website for FOO and its knowledge graphs, offering information on their creation and usage.
(3) An analytical executable notebook to remotely query, visualise and analyse four distributed wildlife knowledge graphs in a granular unified manner.

The proposed interface allows users to script granular SPARQL search queries and to obtain information from remotely located datasets. Our study provides a novel (modular) approach to managing and analyzing wildlife data to support conservation and wildlife management applications. The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 includes the methodology for developing ontology. Section 4 introduces FOODS (Forest Observatory

Ontology Data Store), the proposed ontology-based knowledge graphs. Section 5 evaluates FOODS. Section 6 discusses the proposed system and suggests future work. Section 7 Finally, concludes the paper.

2 RELATED WORK

This section provides an overview of the relevant research on wildlife data management. It begins by examining how Semantic Web technologies are used to model wildlife data and compare different approaches. It then discusses ontologies and knowledge graphs, explaining how they can be developed, and why using ontologies to create knowledge graphs could be beneficial to relevant stakeholders.

2.1 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 [40, 85, 110]. Technologies such as GPS tracking, Wireless Sensor Networks (WSN) [94, 116] and the Internet of Things (IoT) devices [17, 30, 65, 100] collect diverse environmental data and require integration techniques such as knowledge graphs [48, 80, 115]. These graphs provide deep insights into species connections, ecological relationships, and environmental impacts on wildlife [7, 83], thereby supporting interdisciplinary collaboration for effective conservation and decision-making [77, 79].

Past research, such as Athanasiadis et al. [8] developed a semantic framework for large carnivore conservation in northern Greece, integrating animal tracking data with ecological niche modelling for habitat suitability. In contrast, our work differs in location, integration process, output flexibility, and employs an executable interface for interactive analysis across various data types. Wang et al. [112] applied semantic technology to model wildlife observations, including pollution effects on ecosystems, and storing provenance data for traceability. Despite some similarities, our semantic modeling diverges in data transformation using the Resource Description Framework (RDF) Mapping Language (RML) and modular pipelines for scalable data conversion into triple data stores, in contrast to Wang et al.'s manual RDF model conversion.

Mireku et al. [75] and Zheng et al. [96] employ semantic inference for knowledge discovery and predictive analytics to support the dynamics of animal movement trajectories. In contrast, Wannous et al. [113] focused on the development of a trajectory ontology that combines elements from three key concepts, namely, (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 subontologies tailored to specific use cases. In our research, we adopted a different approach by allowing semantically modeled data to populate a foundational ontology. Furthermore, our method includes ontology documentation, publication, and maintenance plans as recommended features.

2.2 Wildlife Ontologies

In computer science, *ontology* is a formal and explicit specification of a conceptualization used to represent knowledge in a particular domain [49]. Ontologies have been used in various domains, including biodiversity, to model knowledge [3]. 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) [91] is an example of a large-scale ontology that covers several domains related to the environment. The Wildlife Ontology (WO) [90] 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 such as 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 [106], formulated as follows:

$$Ontology = (\mathrm{C}, \mathrm{HC}, \mathrm{R}, \mathrm{HR}, \mathrm{I}) \tag{1}$$

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.

2.2.1 Ontology Development Methodologies. We searched the ACM digital library (dl.acm.org) and Google Scholar (scholar.google.com) to identify a suitable methodology. Our search terms included "ontology methodology," "ontology development methodology," and "ontology building approaches." The methodologies we researched include the eXtreme design (XD) methodology [15], which is a modular, incremental approach that maps a set of competency questions to one or more Ontology Design Patterns (ODPs) [42] before integrating them into the ontology under construction. The DILIGENT methodology [86] provides a more flexible trial-and-error approach, recommending the order of discussion, evaluation, justification, and testing in a use-case. METHONTOLOGY methodology [37], 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 needs to be generalized to fit multiple domains. The On-To-Knowledge Methodology (OTKM) [104] focuses on the initial setup, enterprise applications, and maintenance of ontologies. Other well-known methodologies include "Ontology Development 101" by Noy et al. [78] and NeON by Suárez-Figueroa et al. [102]. Whereas the former focuses on ontology conceptualization, 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. [6] and Singh et al. [98]. The Linked Open Terms (LOT) project [87] builds on over two decades of ontological engineering experience, taking inspiration from the Neon methodology [46]. It emphasizes 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 [13, 39] to facilitate their reuse by the research community and software applications. Table 1 compares the different ontology development methodologies.

2.3 Knowledge Graphs

A knowledge graph [14, 51] 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 [31]. Subsequently, the term has been used in various contexts and evaluated by many scholars [20, 21, 55, 81, 118]. A commonly accepted definition of a knowledge graph is one that captures knowledge by defining entities and their relationships [32]. Knowledge graphs offer several benefits in wildlife data management, enabling data integration, unification, linking, and reuse by combining characteristics of different data management paradigms [2].

[1]

Table 1. 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)	[15]	*			*	*	*	*	*			
DILIGENT	[86]		*		*	*	*	*	*			
METHONTOLOGY	[68]	*			*	*	*	*	*			
On-To-Knowledge Methodology	[104]	*			*	*	*	*	*			*
Ontology Development 101	[78]	*			*	*	*	*	*			
NeOn Methodology	[102]	*			*	*	*	*	*			
Linked Open Terms (LOD)	[87]	*	*	*	*	*	*	*	*	*	*	*

2.3.1 Knowledge Graphs Creation Methodologies. A knowledge graph [14, 51] organizes 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 [9]. Another approach involves initiating a knowledge graph without predefining its schema (i.e., ontology) and gradually building both schema and instances during creation. However, designing a knowledge graph schema beforehand can significantly enhance its utility [64]. A six-step process involving data identification, ontology construction, knowledge extraction, data processing, data integration, and knowledge graph evaluation is also commonly used [45]. Furthermore, employing robust tools for linked data, data integration, and data management while continuously analyzing and adjusting deliverables is another viable methodology [10]. The ad hoc creation of knowledge graphs that reuse existing knowledge by interlinking relevant classes and properties from existing ontologies has also been practiced [60]. 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 [74] 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 [27]. 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. [93] and Van Assche et al. [108] contributed to the study area. In addition, Corcho et al. [24] 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.

2.4 Why ontologies for knowledge graphs?

Using ontologies in knowledge graphs reduces ambiguity, ensures data compatibility, and establishes a formal representation of concepts and relationships [16, 58, 67]. 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 [61], enable automated reasoning, and enhance analytical capabilities. Table 2 compares the benefits of building a knowledge graph with and without an ontology.

Table 2. Compares the benefits of building a knowledge graph with ontology and without ontology

Benefits	With Ontology	Without Ontology
Less Ambiguity	Ensures a normalized 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	Enhances targeted search and querying through structured data representation.	Less precise and effective search and querying due to lack of structure.

3 METHODOLOGY

This section proposes the Forest Observatory Ontology (FOO) and outlines the methodology employed for its development. We delve into the entire ontology development life cycle, providing detailed insights into each stage, from the study hub and requirement gathering to the sharing and maintenance phases.

3.1 Dana Girang Field Centre (DGFC)- our research hub

We collaborated with Danau Girang Field Centre (DGFC) (danaugirang.com) as a study hub. DGFC is a research and education facility located in the heart of Sabah, Malaysia, within the Lower Kinabatangan Wildlife Sanctuary. It focuses on the conservation of biodiversity and ecosystems in a region through scientific research. The center studies how wildlife adapts to fragmented landscapes caused by deforestation and human activity. In addition to its research activities, the centre provides educational programs, including internships and field courses for university students, aiming to train the next generation of conservation scientists and increase awareness of environmental issues.

3.2 Forest Observatory Ontology (FOO)

We propose the Forest Observatory Ontology (FOO), a novel upper-level 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 that incorporates classes and properties from established ontologies. FOO standardizes data entities and formalizes their semantics, thereby 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. Additionally, FOO enhances data analysis capabilities by incorporating rules directly into databases to support inferences. To develop FOO, we employed the Linked Open Terms (LOT) methodology [87], 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 1 illustrates the comprehensive lifecycle of FOO's development process.

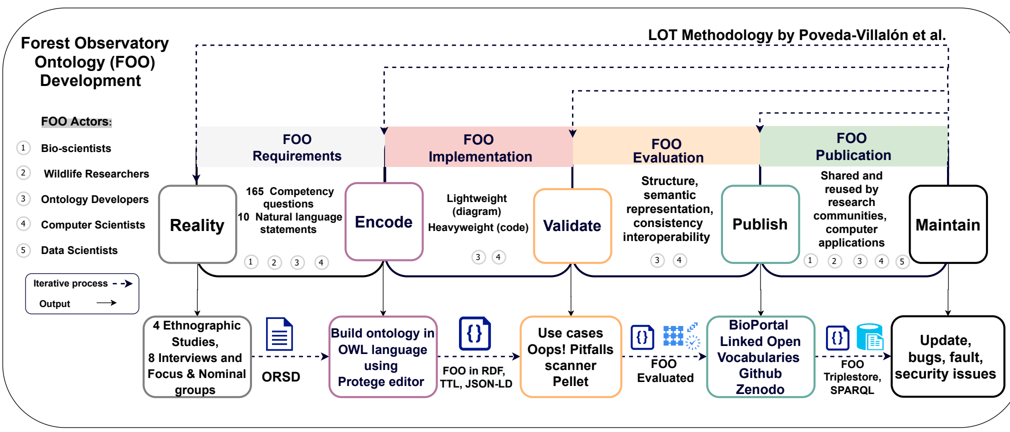


Fig. 1. FOO Ontology Development phases, inspired by Linked Open Terms (LOT) methodology [87]

3.3 Ontology Requirements

During the initial phase of the ontology development process, we created the Ontology Requirements Specification Document (ORSD), adhering to the guidelines outlined in the LOT methodology and as detailed by Suarez-Figueroa et al. [103]. Our ORSD (see Appendix) 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 in identifying use cases for the ontology and selecting the datasets to be modeled.

We compiled a list of Competency Questions (CQs), Natural Language Statements (NLSs), and various use cases for bioscientists and wildlife researchers. CQs, as defined in [50], 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, we 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. We organized a text-based focus group for the third activity, and conducted a nominal group technique session. For each activity, we created three types of administrative documents: 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 for collecting the necessary data was granted by our university's research ethics board. The participants were provided with both online and paper versions of the documents. To enlist participants, we adopted the snowball sampling technique [41] and collaborated with DGFC. We initiated the process with biologists within our network, requesting them to refer to individuals who aligned with the study criteria. Although we initially targeted six participants for formal interviews, our direct engagement strategy proved to be effective, and we successfully conducted interviews with eight participants. In the discussion groups (nominal and focus), we achieved our goal of having at least five participants in each, receiving responses from 14 participants via Google Forms. Interviews were transcribed using Microsoft Word, written notes were taken for the discussion groups and the ethnographic studies were summarized. In line with methodologies adopted in prior research [33], we manually coded the data gleaned from both the interview transcriptions and the notes from discussion groups, coding these insights into Competency Questions (CQs) to guide the ontology's development. After drafting the initial proposal, we collaborated with a domain expert to validate the requirements, as documented in this shared spreadsheet. Table 3 shows the demographic details of the study participants.

No.	Level of Education	Occupation	Years of Experience
1	Master's degree	Student	1 to 3
2		Operational Manager	1 to 3
3		Wildlife Researcher	4 to 10
4	Bachelor's degree	Bio-scientist	Less than 1
5		BSc Student	Less than 1
6		BSc Student	1 to 3
7		Data Manager	11 to 20
8		Wildlife Researcher	Less than 1
9		Bio-scientist	Less than 1
10		BSc Student	Less than 1
11	Doctorate (PhD and DPhil)	Bio-Chemist	1 to 3
12		Bio-scientist	More than 20
13		Bio-scientist	11 to 20
14		Bio-scientist	11 to 20

Table 3. Pseudonymized demographic data of the study participants.

In this table, the entries are grouped by the level of education. This grouping makes it easier to see the distribution of occupations and years of experience within each educational level.

3.3.1 ***Ethnography***. We conducted ethnographic research at DGFC in the summer of 2022 to observe the collection and processing of wildlife data (Figure 2) [69]. Four activities were carried out: (i) comparing butterfly diversity, (ii) comparing Proboscis monkey activity, (iii) finding the tracked Sunda pangolin, and (iv) finding the Elephas maximus (Asian elephants).

(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.

Fig. 2. Forest Observatory Ontology Development's activities collage

(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 behavior 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 h of recordings from each device, are yet to be analyzed 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, we entered the forest, equipped for protection against insects and rain, aiming to locate the Sunda pangolin using a noise-emitting antenna designed for close 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): We left from Sandakan Jetty, heading towards the DGFC center through the dense forest. On our boat ride, we suddenly saw a group of Asian elephants swimming in the lower Kinabatangan River.

3.3.2 ***Interviews***. We carried out eight semi-structured via face-to-face discussions with specialists in genetics and biology focusing on wildlife conservation. Seven of the interviews were based in Sabah (Malaysian Borneo), apart from one individual from the United Kingdom who volunteered in the DGFC. The participants had a diverse range of experience in landscape ecology and conservation biology research, with their experience spanning from one to twenty-five years. To recruit participants for the study, we identified the first bioscientist from our university mailing list, then we employed a snowball sampling technique [1], in which each participant recommended another potential participant. We then provided these nominees with an information sheet about the interview and details of ethics approval two weeks prior to their interviews. During these sessions, we followed a consistent semi-structured guide to delve into the types of data the participants collected and processed as well as their aims to use these data to make well-informed decisions swiftly. Every participant completed their interview within 60 minutes, and we preserved the audio recordings for detailed analysis. The interview questions covered a range of topics, including:

- What is your opinion about a given User Interfaces mock-ups?
- 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 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)?

3.3.3 Interviews analysis and findings. Our analysis of the interview transcriptions was conducted using inductive coding [12]. This approach entails a thorough examination of the data, including interview transcripts, field notes, and documents, to identify text segments of interest or significance. Each segment was labeled 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 a cyclical process, we kept comparing new data to existing codes and themes, refining them as needed. Table 4 presents the themes that emerged from this process along with their descriptions. The research findings revealed a collective desire for improved data management, visualization, and accessibility across different 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, the integration of disparate data sources, and the need for better tools to visualize and analyze data, particularly 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 labeled the pictures in the timestamp of the name and using the name.*" showing efforts to improve data organization 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 visualize current data, such as GPS tracking, while adapting to evolving conservation needs by incorporating new data types and analytical methods.

3.3.4 ***Focus and nominal groups.*** In the form of visual materials, we created a map of Sabah, Malaysian Borneo, 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. Attendees were provided with a copy of the primary map and three data type cards and were requested to suggest ideas, potential use cases, and questions that could be addressed using these different datasets. From the discussion groups, we collected a list of use cases for FOO and the relevant datasets, exploring their potential applications and usefulness in informed decision-making. Screenshots of the maps and cards with written information are included in the appendix. This exercise has been a great portion of CQs and NLSs for ontology, offering valuable insights into its development and application. Participants gathered various perspectives and

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 visualization tools, search functionality, or customization 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 the transformation of 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 the use of databases, APIs, or interactive dashboards.
Challenges	Identified issues with current data systems, such as 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' conceptualization of the ideal data model or system.

Table 4. Overview of Participant Feedback Themes. This table outlines the main feedback themes from the evaluation of our proposed data management system, including design impressions, functional requirements, and user experiences.

Table 5. Functional Requirements Validation Criteria as stated by [50]

Criteria	Results	Bio-scientists	Wildlife Researchers	Ontology Developers	Computer Scientists
Correctness	All requirements relate to FOO's concepts.	✓	✓		
Complete	The intended user confirmed FOO's sufficiency.	✓	✓		✓
Consistent	There were no conflicts between FOO requirements.	✓	✓	✓	
Clear	Each requirement has one precise meaning.	✓	✓	✓	✓
Concise	All requirements were relevant.	✓	✓	✓	✓
Comprehensible	The stakeholders understood FOO requirements.	✓	✓		✓

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 development sheet containing the CQs and NLSs, along with the ontology requirements specification document (ORSD), were finalized and uploaded to the ontology GitHub repository.

3.3.5 ***Ontology requirements validation***. Following the methodology proposed by Grüninger and Fox [50], functional requirements must meet certain standards prior to formal acceptance. We evaluated the functional requirements of the ontology against these criteria, as outlined in Table 5. The stakeholders involved in the FOO project confirmed that these requirements were (i) accurate, (ii) comprehensive, (iii) coherent, (iv) unambiguous, (v) concise, and (vi) clearly defined.

Table 6. Lists and compares IoT and wildlife ontologies. SSN = Semantic Sensor Network [23], SOSA = Sensors, Observations, Samples, and Actuators [59], S3N = Smart Sensor Network [95], SSxN [23], IoT-Lite, IOT-o [97], OBOE = Extensible Observation Ontology [70], SAREF = The Smart Applications REFerence [26], ncbitaxon, Ecocore , BBC-WO = BBC wildlife Ontology, African-Wo = African Wildlife Ontology [62], GeoSpecies ontology .

	IoT Domain								Wildlife Domain				
FOO CLASS	SSN	SOSA	SSNX	IoT-LITE	SAREF	OBOE	S3N	IOT-O	ncbitaxon	ecocore	BBC-WO	African-WO	GeoSpecies
Sensor	*	*	*	*				*					
Observation	*	*	*			*		*					
Observable Property	*	*					*	*					
Feature of Interest	*	*			*		*	*					
Results	*	*	*	*			*						
ResultsTime	*	*	*			*		*					
TaxonRank											*		*
TaxonName											*		
Person											*		
Image											*		

3.4 Ontology Implementation

Based on the Ontology Requirements Specification Document (ORSD) crafted in the requirements phase, we established that the FOO's scope would include both Internet of Things (IoT) concepts, such as sensor and its observation, and wildlife aspects, like animals. For example, our 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 geographic location observations at different and equally spaced time intervals along with temperature readings at each specified interval. We conducted a search across various scholarly resources and ontology repositories to identify ontologies relevant to our research. The search included Google Scholar [47], BioPortal repository, and other pertinent websites. Our selection criteria stipulated that publications must be published between 2015 and 2020. We used a variety of search terms, such as "sensor data ontology," "semantic modeling for sensor data," "semantic IoT data," and "IoT ontology."

We found several domain-specific ontologies for modeling sensors and wildlife data. The Semantic Sensor Network (SSN) ontology describes the sensory observation processes (SSN, SSN2). Within SSN Version 2, there is a Sensor, Observation, Sample, and Actuation (SOSA) ontology that is suitable for lighter use without the full SSN [59]. IoT-Lite ontology provides foundational descriptions of IoT resources, while the Smart Applications REFerence (SAREF) ontology focuses on referencing IoT appliances [26]. The Extensible Observation Ontology (OBOE) [70] 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 [63], 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, we manually filtered out the most commonly used ontologies for modeling sensor data, such as the SSN [111]. Among the shortlisted ontologies, the SAREF ontology [26], which is designed for smart appliances, IoT devices, and services; however, it may not adequately model sensor data observations. IoT-Lite ontology [11] provides a basic framework of classes and properties for describing IoT devices, sensors, and actuators. However, for our specific use cases, we needed more classes 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 customized to fit different use cases, allowing for interoperability across various IoT systems and domains. Although it covers different 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 modeling sensor observations. The FIESTA-IoT ontology (iot.ee.surrey.ac.uk/ontology/fiesta-iot.owl) primarily focuses on modeling IoT-related concepts but includes more entities than we need. It incorporates classes from the SSN ontology (Version 1)[23], 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.

After conducting an in-depth examination and comparison of contemporary ontologies, we adopted concepts from SSN ontology (Version 2) [111]. 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 [59], and the Quantities, Units, Dimensions, and Types (QUDT) ontology. Such integration makes the SSN ontology (Version 2) well-suited to our needs. FOO applied the extraction of Ontology Design Patterns (ODPs) as a way of ontology reuse [38, 88]. FOO extracted SOSA ontology from SSN version 2 using owl:imports. FOO classes ‘foo:Sensor‘, ‘foo:Observation‘, ‘foo:ObservableProperty‘, and ‘foo:FeatureOfInterest‘ were aligned with the corresponding SOSA classes (‘sosa:Sensor‘, ‘sosa:Observation‘, ‘sosa:ObservableProperty‘, and ‘sosa:FeatureOfInterest‘) using ‘owl:sameAs‘. 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/). As elaborated in Section 2.2, our analysis encompasses various ontologies relevant to wildlife concepts. Our modelling efforts concentrated on sensor data observations, geographical locations, data units, and wildlife characteristics, as shown in Table 6.

Sensor data observations: We adopted classes and properties from the SOSA ontology.

Location: We relied on GPS coordinates, as this work focuses on outdoor locations.

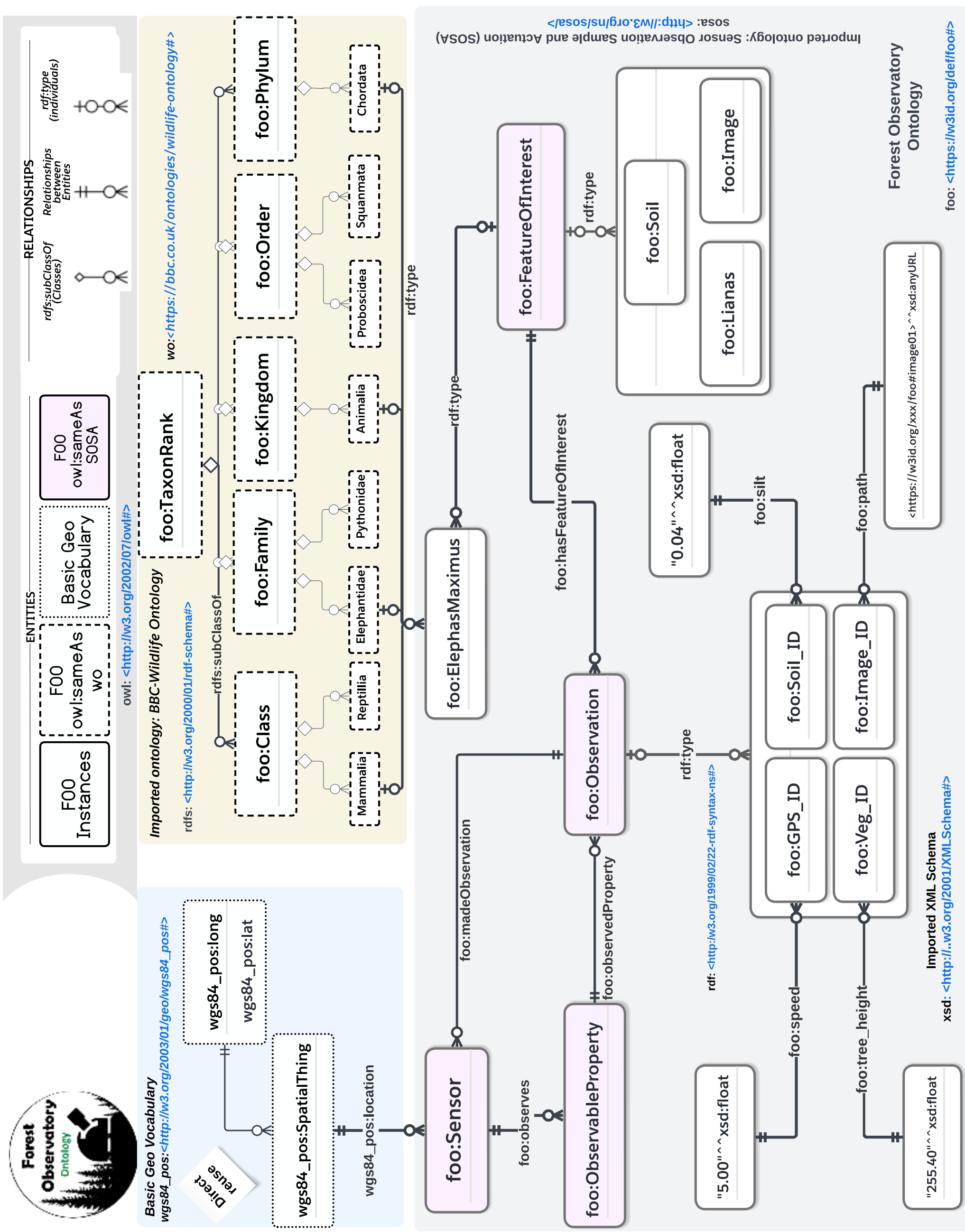
Temporal aspects: A timestamp is a crucial element in semantic modeling. This enabled us to differentiate observations within the same dataset and link them to those in other datasets. For this project, we opted to record the timestamp of each observation using the XML DateTime data type (xsd:dateTime). Although OWL-Time ontology can be used to model dates and times, the SPARQL Time function fulfills a similar role. Consequently, we decided to omit the OWL-Time from FOO.

Units of data: To model the units of our observations, we chose to reuse classes from the Quantities, Units, Dimensions, and Types Ontology (QUDT) [101], which is part of the SSN (version 2) ontology.

Wildlife features: We explored several wildlife ontologies (see Table 6), including the African Wildlife ontology [62] and the BBC Wildlife Ontology. Ultimately, we selected the BBC Wildlife Ontology (WO) for its comprehensive coverage of concepts and properties, such as the hierarchy of taxonomic ranks encompassing all levels of biological classification. Figure **??** shows the classes reused from the SOSA and BBC wildlife ontology and their relationships, sharing the common upper class (owl:Thing).

Hence, we created and discussed conceptual models (i.e., diagrams) with ontology stakeholders. Following the graphical representations, we encoded FOO in Web Ontology Language (OWL), edited it with Protégé (protege.stanford.edu), and developed pipeline codes in Python to serialise the datasets populating FOO. We publish and maintain all data and ontological resources in dedicated GitHub repositories and a website. Figure 3 illustrates the design of the proposed ontology... Table

7 summarizes 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 and



)

Fig. 3. Lightweight version of the Forest Observatory Ontology (FOO), main classes, properties and instances.

Table 7. Main Classes and relationships from the The Forest Observatory Ontology (FOO), foo= https://w3id.org/def/foo#

OWL Class	Preferred Label	Description
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. Male proboscis monkeys have notably large noses, which are believed to have evolved due to their sexually competitive social system. Significant correlations exist between nose size, body size, testis size, and the number of females in a male's harem. This suggests that both male competition and female choice have driven the evolution of these enlarged noses.
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. Despite its high protection status in Sabah, where it is (totally protected) under the Wildlife Conservation Enactment 1997, the species faces severe threats from illegal wildlife trade and habitat degradation. In 2019, authorities in Sabah seized over 30 tonnes of pangolin products, highlighting the extensive illicit trade network. The Sunda pangolin occupies various habitats, from primary and secondary forests to wetlands, mangroves, and grasslands.
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	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	Observable Property	An observable quality (property, characteristic) of a FeatureOfInterest. (e.g., Temperature, humidity, presence)
foo:Observation	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	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).

3.5 Ontology Evaluation

We investigated various ontology evaluation techniques and discovered that ontology evaluation primarily focused on assessing quality and accuracy during and after its development [82]. Raad et al. [89] identified four ontology assessment methods from the literature: (i) gold standards, (ii) corpus-based, (iii) criteria-based, and (iv) task-based. McDaniel et al. [72] described ontology evaluation as a two-fold process, namely, the glass-box and black-box approaches. The former evaluates the ontology incrementally throughout its lifecycle, which is also known as component evaluation. By contrast, the latter is a task-based approach associated with evaluating an ontology's performance in a specific task or application [73]. The most suitable method for evaluating an ontology depends on the intended purpose. Our proposed ontology, FOO, is designed to support applications that integrate heterogeneous data sources for decision-making. Thus, we initially evaluated *structure, semantic representation, and interoperability*. To assess the structure and semantic representation, we selected the open-source online scanner, Oops! (oops.linkeddata.es), as the baseline, and the domain expert feedback for evaluating semantic representation. Subsequently, we applied the black-box (i.e., task-based) approach to assess the applicability and interoperability of knowledge graphs (i.e., FOO instantiated with heterogeneous RDF datasets), focusing on how well it addresses use cases and their efficiency in data exchange between different computer systems. The finalized details regarding the validation criteria after implementation are listed in Table 9.

3.5.1 FOOPS! Evaluation. FOOPS! [44] 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. For the **Findable** dimension, it conducts nine checks that assess whether the ontology URI is persistent and resolvable, includes a resolvable version IRI that is unique for each version and contains minimum descriptive metadata such as title and description. Moreover, it verifies if the ontology prefix and namespace are registered in external registries like prefix.cc (prefix.cc/foo) and Linked Open Vocabularies (LOV) (lov.linkeddata.es/). In the **Accessible** dimension, three checks ensure proper content negotiation with at least one RDF serialization and HTML format and verify that the URI protocol is open 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, in the **Reusable** dimension, nine checks verify the availability of human-readable documentation, ensure the presence of provenance metadata, license information, and detailed vocabulary metadata, and check that ontology terms are well-described with labels and definitions. Our ontology FOO passed the test and scored a respected 78%, outperforming SOSA ontology with a score of 67%. Figure 4 shows a screenshot of the test results of both SOSA and FOO.

3.5.2 Evaluation with SPARQL Queries. To evaluate FOO with SPARQL queries, We first needed to understand the structure of the ontology by inspecting its classes, properties, and instances. We then formulated SPARQL queries to explore FOO and subsequently evaluated the performance of each query. The table presented in 8 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 2, and recall 3. 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

Fig. 4. FOOPS pitfall scanner results of our ontology FOO compared to SOSA ontology.

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 8. 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:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

Table 9. FOO's Validation after Implementation

Criteria	Results	FOOP!	Pellet [99]	Bio-Scientists
Consistent	There were no conflicts between FOO requirements.	✓	✓	✓
Verifiable	FOO was able to answer complex questions.			✓
Clear	Each entity has one precise meaning.	✓		✓
Comprehensible	The stakeholders understood FOO requirements.			✓

The recall metric is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

3.5.3 FOO Validation. We returned to our participants with the FOO documentation, themes, CQs that FOO is expected to answer. We also provided them with access to an interactive dashboard to visualize elephant movements. Table 9 summarizes the evaluation outcomes.

3.6 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 prioritize 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 [25].

We selected WIZARD for DOCumenting Ontology (WIDOCO), a tool based on the Live OWL Documentation Environment (LODE), which is utilized within the seven-star linked data model platform [43, 84], to document FOO. WIDOCO enabled us to generate HTML pages that present human-readable and machine-readable visualizations of FOO along with Oops! evaluation. Moreover, we used OnToology (ontoology.linkeddata.es) [5] to secure a persistent identifier for FOO documentation (https://w3id.org/xxx/xxx#), ensuring it can be reliably referenced and accessed over time under the Creative Commons 4.0 International SA (CC BY-SA 4.0) license. We 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, we 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 websites. Figure 5 conveys screenshots from the websites.

4 FOODS

This section describes our approach to modelling wildlife datasets using semantic data modelling. First, by leveraging FOO, we constructed four distinct knowledge graphs from the datasets of interest. Next, we demonstrate how to connect and query knowledge graphs (i.e., data sources) on a single display. This approach enabled the representation and integration of diverse wildlife data sources.

4.1 Overview

We developed FOO to create four distinct wildlife knowledge graphs for a wildlife research facility. Figure 6 shows the relationships between the proposed ontology (FOO) and wildlife knowledge

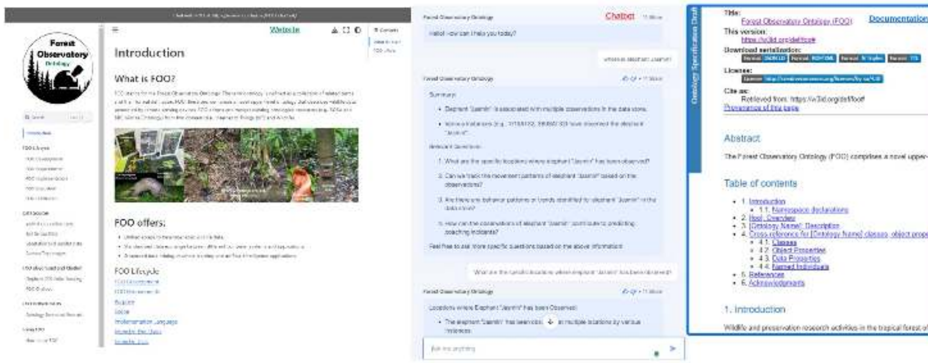


Fig. 5. The screenshots show FOO's interactive website and documentation. They feature interactive HTML elements that detail the development process and guidelines for usage, along with links to downloadable datasets and the GitHub repository. The middle screenshot is a conservation Artificial Intelligence (AI) chatbot embedded in FOO's website to ease question-answering and granular data acquisition. The documentation contains entities' definitions and links to the ontology file, downloadable in multiple formats.

graphs. To transform four wildlife datasets—encompassing soil data, vegetation and site habitats, GPS collar data, and trail camera images into knowledge graphs, we used the Matey web user interface (rml.io/yarrrml/matey), powered by YARRRML (Yet Another RDF Rules Language) [54, 109]. YARRRML (rml.io/yarrrml) specifies a set of prefixes to create namespaces and offers mapping rules to generate RDF triples from the data sources. In addition, we developed modular pipelines to manage large data volumes, ensuring data serialization with names or schemas that align with those defined in FOO. Figure 6 illustrates the connections between FOO and wildlife datasets. Resources related to this study, including the ontology and code, are available on the proposed website.

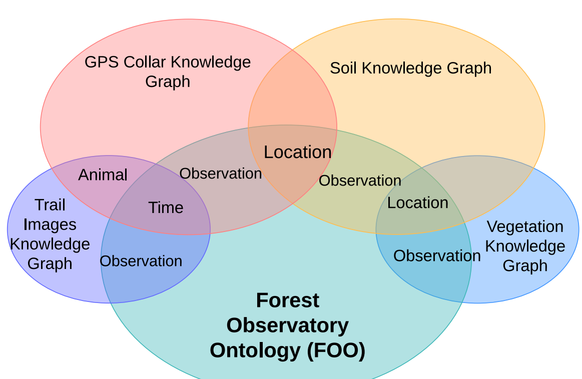


Fig. 6. Main related concepts between FOO and the proposed knowledge graphs

4.2 Soil Knowledge Graph

Based on the experience drawn from ontology development, we decided to outsource the soil data. The selected 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) [34]. 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 1, 2 defined the rules for the "Observation" entity. They 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.

Listing 1. Soil data prefixes

```
foo:  "http://w3id.org/xxx/xxx#"
xsd:  "http://w3.org/2001/XMLSchema#"
sosa:  "http://w3.org/ns/sosa/"
```

Listing 2. Soil data YARRRML

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

Figure 7 shows the classes and instances distribution for the soil knowledge graph. Meanwhile, Table 10 describes them. Table 11 provides a descriptive analysis of the modelled data.

Table 10. Soil data set variables Description

Name	Instance of Class	Data Type	Description
foo:Identifier	foo:Observation	rdf:type	Unique Sample Identifier
foo:site	foo:ObservableProperty	xsd:string	Geographical area/site which samples were taken from
foo:land_Use	foo:ObservableProperty	xsd:string	Land use of the study plots: Unlogged tropical forest, Logged tropical forest or Oil palm plantation
foo:plot_Name	foo:ObservableProperty	xsd:string	Name of the 1 Ha plot sampled
foo:subplot	foo:ObservableProperty	xsd:string	Number of the subplot sampled within each 1 Ha plot
foo:horizon	foo:ObservableProperty	xsd:string	Soil horizon sampled
Soil_Moisture	foo:ObservableProperty	xsd:float	Gravimetric soil moisture
foo:horizon_Depth	foo:ObservableProperty	xsd:float	Depth of the organic soil horizon sampled
foo:bulk_Density	foo:ObservableProperty	xsd:float	Measured Bulk Density of soil sample
foo:soil_pH	foo:ObservableProperty	xsd:float	Measured pH of the soil sample
foo:total_C	foo:ObservableProperty	xsd:float	Total carbon content of the soil sample
foo:total_N	foo:ObservableProperty	xsd:float	Total nitrogen content of the soil sample
foo:inorganic_P	foo:ObservableProperty	xsd:float	Inorganic/soluble phosphorus concentration of the soil sample
foo:C:N	foo:ObservableProperty	xsd:float	Carbon to nitrogen ratio of the soil sample
foo:C:P	foo:ObservableProperty	xsd:float	Carbon to inorganic phosphorus ratio of the soil sample

4.3 Vegetation Knowledge Graph

These datasets 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 encompassed plant properties, measures of forest structure, and metrics of forest fragmentation 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 [35]. Figures 8 and Table 12 illustrate

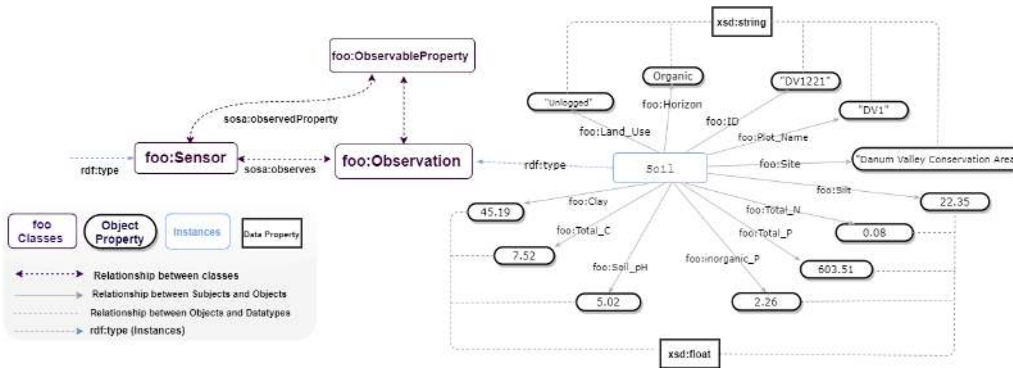


Fig. 7. Soil Knowledge Graph

Table 11. Soil data descriptive analysis

	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
Mean	26.71	3.74	0.69	5.59	6.04	0.39	33.97	15.04	0.27
STD	9.37	1.91	0.28	0.85	4.61	0.21	51.56	3.52	0.16
MIN	7.62	0.20	0.17	3.22	0.83	0.09	3.77	6.65	0.01
25%	20.56	2.35	0.49	4.92	3.74	0.27	14.34	12.96	0.17
50%	26.43	3.50	0.66	5.58	4.94	0.34	20.49	14.46	0.24
75%	31.91	5.00	0.86	6.32	6.33	0.43	32.91	16.47	0.33
MAX	65.10	9.50	1.84	7.42	33.45	1.49	571.25	39.59	0.89

the vegetation knowledge graph modeling and its descriptive data analysis, respectively. The RDF mappings were modeled 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, 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.

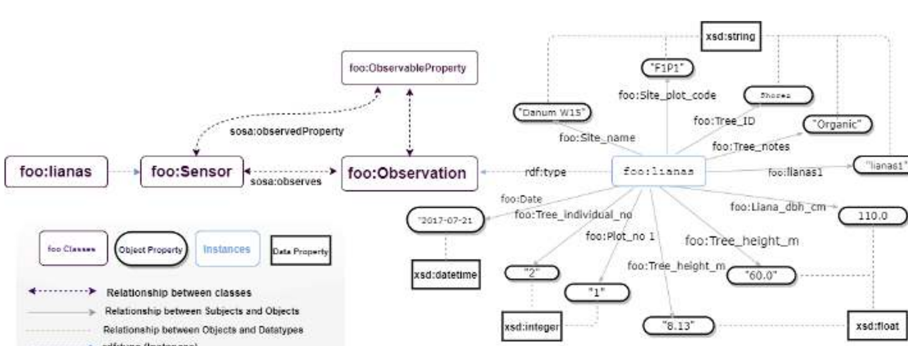


Fig. 8. Vegetation Knowledge Graph

Table 12. Lianas data descriptive analysis

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

Table 13. GPS Collar Sensor Metrics

Name	Instance of Class	Data Type	Description
foo:ID	foo:Observation	rdf:type	A unique identifier for each GPS collar sensor observation.
foo:localDate	foo:ObservableProperty	xsd:date	The local date in Sabah, Malaysia, when the GPS collar records its readings.
foo:localTime	foo:ObservableProperty	xsd:time	The local time in Sabah, Malaysia, when the GPS collar records its readings.
foo:gMTDate	foo:ObservableProperty	xsd:date	The date in GMT for standardising time across data collections.
foo:gMTTime	foo:ObservableProperty	xsd:time	The time in GMT for standardising time across data collections.
pos:lat	wgs84_pos	xsd:float	Latitudinal coordinate of the elephant at the moment of data collection.
pos:long	wgs84_pos	xsd:float	Longitudinal coordinate of the elephant at the moment of data collection.
pos:alt	wgs84_pos	xsd:float	Altitude of the elephant in meters at the moment of data collection.
foo:temperature	foo:ObservableProperty	xsd:double	Estimated temperature of the elephant in Celsius at the moment of data collection.
foo:speed	foo:ObservableProperty	xsd:float	Speed of the elephant at the moment of data collection.
foo:direction	foo:ObservableProperty	xsd:float	Direction of elephant travel at the moment of data collection.
foo:distance	foo:ObservableProperty	xsd:float	Distance (m) travelled from the last to the current data collection point.
foo:count	foo:ObservableProperty	xsd:integer	Observation count per data set.
foo:hdop	foo:ObservableProperty	xsd:integer	Horizontal Dilution of Precision (HDOP), indicating GPS accuracy in latitude and longitude. Lower values indicate better precision.

4.4 GPS Collar Knowledge 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 13 [36, 69]. Figure 9 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, prioritizing the protection of these endangered species [36, 69]. In addition, Figure 10 exemplifies how a sensor's observations (GPS collar named after elephant Abaw) were connected to SOSA and BBC-wo ontologies.

4.5 Camera Trap Images Knowledge Graph

A dataset containing 1000 images of Asian elephants was modeled. Prior to their transformation into RDF graphs, the metadata of the images were extracted and stored as CSV files. The RDF dataset includes unique paths that point to image locations on a protected cloud server. Figure 11 and Table 14 illustrate the results of the semantic modeling and the data entities, respectively.

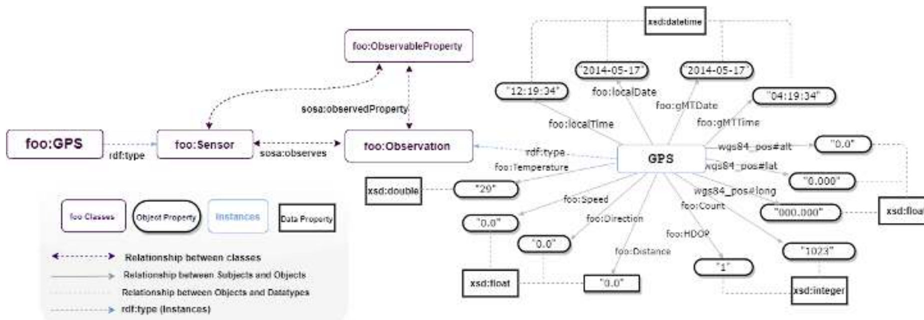


Fig. 9. GPS Collar knowledge graph

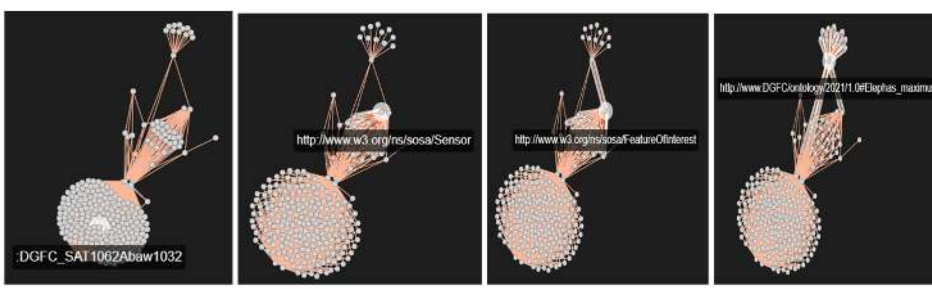


Fig. 10. Integrating the GPS collar data with other components involves a set of blocks. The initial block corresponds to the data observations, followed by the second block, which links the SOSA sensor via an rdf:type observation. The third block connects the interest category feature to the Elephas Maximus (Asian elephant) instance type. Finally, the fourth block combines the BBC wildlife ontology by formally defining Elephas maximus as an elephant.

Table 14. Trail Camera's metrics

Namespace	Instance of Class	Data Type	Description
foo:DGFC_ID	foo:Observation	rdf:type	Unique Sample Identifier
foo:name	DGFC:Name	xsd:string	The named assigned to an image at collection time
foo:path	foo:ObservableProperty	xsd:string	The URI to point at the location of the image in secure cloud
foo:localDate	foo:ObservableProperty	xsd:date	The current local date in Sabah, Malaysia when the GPS collar collects its readings.
foo:localTime	foo:ObservableProperty	xsd:time	The current local time in Sabah, Malaysia when the image collects its readings.
foo:model	foo:ObservableProperty	xsd:string	The model of the trail camera used to capture the image
foo:make	foo:ObservableProperty	xsd:string	The make of the trail camera used to capture the image

4.6 FOODS

The overall architecture and elements of FOODS are depicted in Figure 12. 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, along with the Forest Observatory Ontology (FOO), are stored together in a unified database known as a triple store.

The process of 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 [107]. The role of the ontology is critical as it serves to interlink these diverse datasets. It

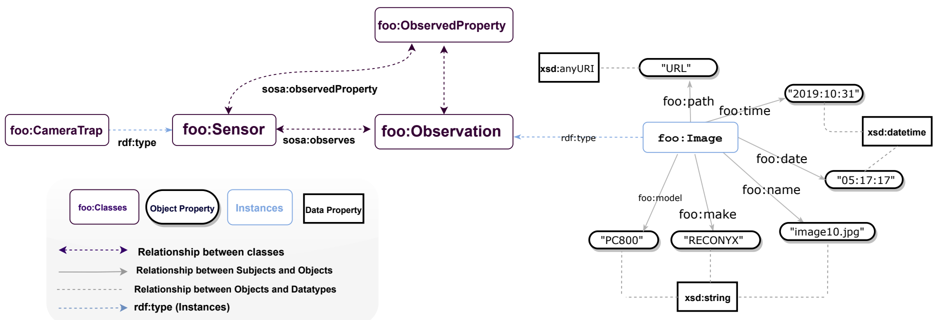


Fig. 11. Camera Trap images knowledge graph conceptual model

achieves this by using a common URL for each dataset, which referenced their conversion into RDF format. Once these RDF formatted datasets are merged within the same triple store, they inherently connect and form a unified graph. This unified graph not only maintains data integrity but also enhances data interoperability across different research datasets, thereby creating a comprehensive resource for ecological research.

Furthermore, the knowledge graphs are not only stored but also published as URLs, which can be accessed and parsed using libraries such as RDFLib in Python environments like Colab. This accessibility allows for easy integration with existing data analysis tools and enhances collaborative opportunities across the scientific community. To interact with and use these knowledge graphs, we employ the SPARQL Protocol and RDF Query Language [4], which facilitates various operations such as data mapping, virtualization, and interactive visualization. This query language enables authorized users— including wildlife researchers, data scientists, and developers— to access and manipulate the graphs.

FOODS provide a powerful foundation for integrating Artificial Intelligence (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 be used to extract entities, relationships, and attributes from unstructured text and populate the knowledge graph, while 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 Artificial Intelligence (AI) models.

The flexible knowledge representation in the graph also enables the use of 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 Artificial Intelligence (AI) systems more transparent and explainable by tracing the reasoning behind outputs using the encoded relationships and logical rules. Knowledge graphs can integrate diverse data sources, and techniques like entity resolution and data fusion can be applied to maintain data quality and consistency. The synergistic combination of FOODS and Artificial Intelligence (AI) technologies is a key driver of the growing field of "Contextual AI", enabling more intelligent, contextual, and explainable systems.

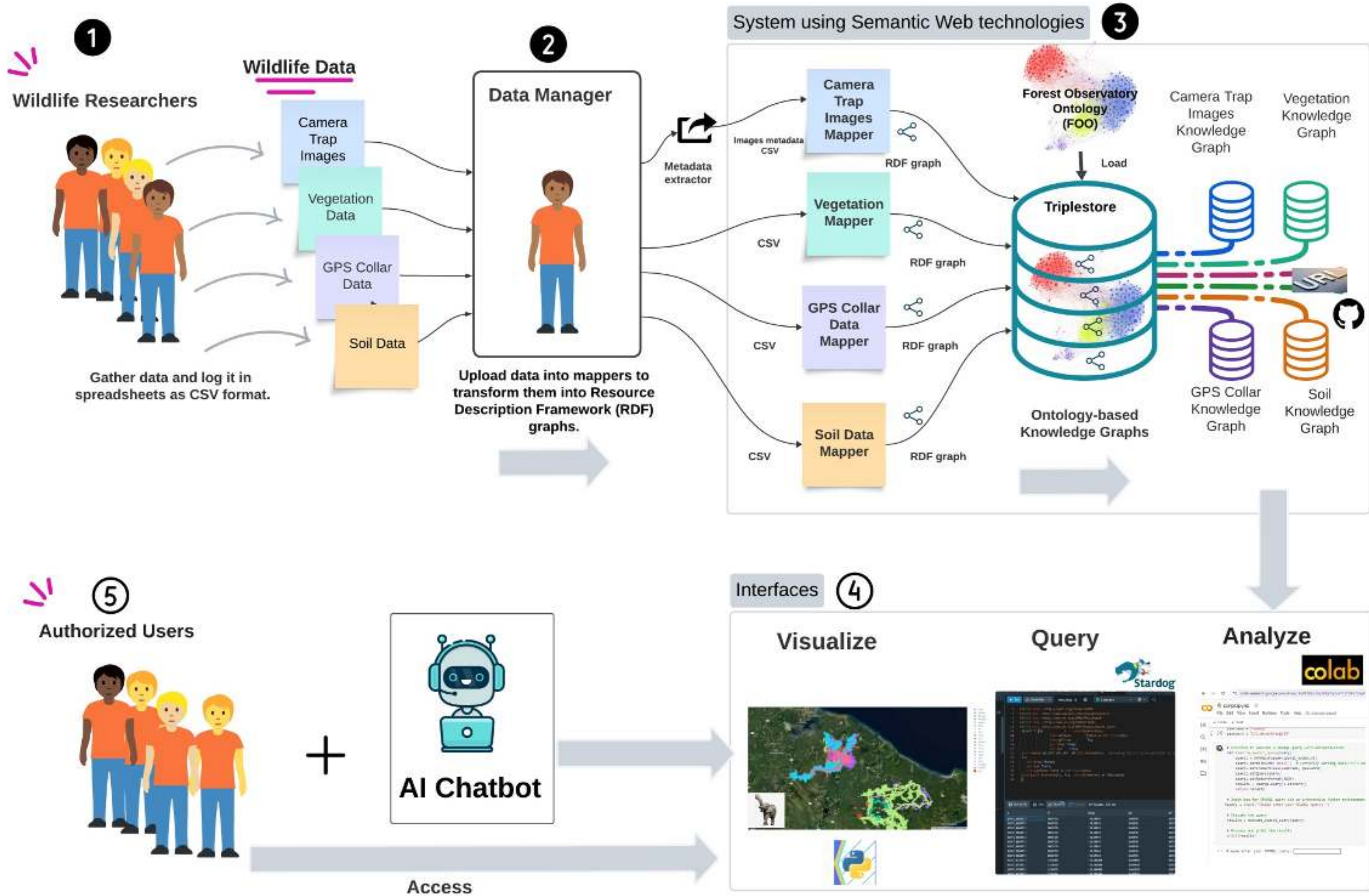


Fig. 12. This figure shows the management of wildlife research data through RDF graph assignment by a data manager, with RDF graphs and the ontology FOO stored in a unified database (triple Store). It highlights the process of knowledge graph generation as a mapping between source and target schemas, focusing on RDF for the target schema. The division of FOODS 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 15. 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 visualization of different animals' movements. Q9. I am satisfied with how the data integration tool handles complex data sets.

5 FOODS EVALUATION

This section assesses FOODS using a task-based approach. To evaluate the usefulness of our proposed tool, we conducted a usability study involving nine domain experts to assess system performance; six of these experts are related to DGFC and had participated in the discussion groups during the ontology requirement gathering phase. We then discuss three use cases—5.2, 5.3, and 5.4—derived from the requirements outlined in Section 3.3. We conducted an in-depth evaluation of the third use case (5.4), as it encapsulates a real-life scenario. This evaluation highlighted the primary benefits of FOODS, particularly through the extraction of several competency questions from use cases.

5.1 Domain Experts Evaluation

The usability study for testing FOODS was judged by the presence of a conservation biologist among the participants. Participants were provided presentations about the dashboard and how to query and analyze the knowledge graphs. Responses were quantified on a Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) across several aspects such confidence in the tool, its usefulness, ease of learning, performance, time efficiency, UI clarity, need for technical support, visualization quality, and data handling satisfaction. Table 15 shows participant feedback on our proposed FOODS across various dimensions. We analyzed the results and visualized them, reducing the dimensions to confidence, ease of use, learning curve, performance, and data handling capabilities for simplicity (Figure 13).

5.2 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 traveling, foraging, and socializing. Their

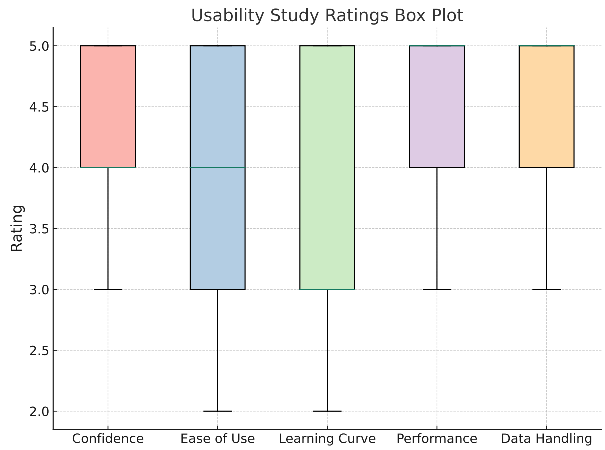


Fig. 13. Box plot representation of usability ratings across five key metrics (Confidence, Ease of Use, Learning Curve, Performance, Data Handling) as evaluated by participants in a usability study. Each box indicates the interquartile range (IQR) with the median highlighted, encapsulating the central tendency and dispersion of ratings, thus providing insights into the tool's perceived effectiveness and user satisfaction.

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 behavior 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 to identify key habitats for conservation, reduces human-elephant conflicts, and guides the allocation of resources such as deploying rangers and fitting motion-activated cameras. These patterns reveal the areas where elephants move and find essential resources, allowing for focused conservation initiatives, strategies to minimize conflicts, and effective usage of anti-poaching resources. Moreover, it enables researchers to deduce when different elephants spend time together. For example, observing two elephants traveling 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, we formulate the SPARQL query in 3. It shows how to select *observations* of certain elephants based on a geospatial criterion—specifically, identifying observations within a 5-kilometer radius of the Lower Kinabatangan Wildlife Sanctuary on January 1, 2023. This query is structured to work within the constraints of SPARQL's capabilities, assuming the absence of direct support for complex geospatial functions or extensions, such as GeoSPARQL.

5.3 Use Case 2: Salt licks locations

Salt licks, also known as 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, harbor 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, which can harm 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 (4) 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 name of the elephant, which can be selected in the query.

Listing 3: Retrieve observations of elephants (?elephantGPS) and images (?image) recorded on January 1st, 2023, within a 5-kilometer radius of the Lower Kinabatangan Wildlife Sanctuary.

```
SELECT * {
?elephantGPS a sosa:Observation;
            foo:localDate "2023-01-01"^^xsd:date;
            pos:long ?long;
            pos:lat ?lat.
  ?image a wo:Image;
            foo:localDate ?Date.
# Estimate location of the Lower Kinabatangan Wildlife Sanctuary
  BIND(5.40 AS ?targetLat) # Target latitude
  BIND(118.08 AS ?targetLong) # Target longitude
# Calculate latitude and longitude bounds for a 5km radius
# Note: These are rough approximations
  BIND(?targetLat - (5/111) AS ?minLat)
  BIND(?targetLat + (5/111) AS ?maxLat)
  BIND(?targetLong - (5/(111 * COS(RADIANS(?targetLat)))) AS ?minLong)
  BIND(?targetLong + (5/(111 * COS(RADIANS(?targetLat)))) AS ?maxLong)
  FILTER (
    ?lat >= ?minLat && ?lat <= ?maxLat &&
    ?long >= ?minLong && ?long <= ?maxLong )}
```

Listing 4: Assuming that the Danum Valley has salt licks and we want to formulate a SPARQL query to select elephants observed at the Danum Valley and gather information about the salt licks in the area.

```
SELECT ?elephant ?elephantGPS ?saltLick ?saltLickLocation ?saltLickArea {
  # Select observations of elephants with GPS coordinates
  ?elephantGPS a sosa:Observation;
               foo:localDate ?date;
               pos:long ?elephantLong;
               pos:lat ?elephantLat.

# Filter observations to those within the Danum Valley
  FILTER (?elephantLat >= DanumValleyMinLat && ?elephantLat <= DanumValleyMaxLat &&
          ?elephantLong >= DanumValleyMinLong && ?elephantLong <= DanumValleyMaxLong)

  # Select salt licks and their locations in the Danum Valley
  ?saltLick a saltlick:SaltLick;
            saltlick:location ?saltLickLocation;
            saltlick:area ?saltLickArea.}
```

5.4 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 visualize 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 the elephant was injured by a snare near an oil palm plantation. The officer promptly notified the manager, who then 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. Currently, our system uses historical data due to availability. However, at the reproduction level, our system can stream real-time data from sensors into our pipeline codes for transformation into RDF.

We developed such use cases to evaluate the proposed knowledge graphs by 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 performance and effectiveness of the knowledge graphs in achieving these objectives. We derived four competency questions based on the third use case, Listing (5.4), which depicts a real-life scenario. These questions were formulated 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, and how close was it?
CQ3: What are the soil metrics near the elephant Jasmin?
CQ4: What are the other elephants near the palm oil plantations?

The first query in Listing (5), investigates how GPS collar information can be merged with camera-trap images to authenticate elephant identities and assess the urgency of incidents. Listing (6) aims to elucidate elephant behaviors, such as foraging and socializing. Listing (7) 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. Listing (8) leverages the reasoning capabilities of Semantic Web technologies to introduce assertive rules into data. For instance, a logical rule might stipulate that if a snare injures one elephant, the other nearby elephants could also be at risk. We used Federated SPARQL queries to interrogate the knowledge graphs, enabling us to retrieve answers from any integrated data source within the FOO. Responses were obtained from relevant knowledge graphs, including those containing data from the GPS collar and soil datasets incorporated into FOO. Sample SPARQL queries are provided, and a comprehensive description of the competency questions, queries, and their corresponding answers can be accessed on the ontology website and its GitHub repository.

Listing (5) **What are Jasmin's observations between 2012-02-07 to 2012-02-15?**

```
SELECT *
{ GRAPH <urn:Jasmin> {
?s DateTime:gMTDate ?date;
foo:Direction ?JasminDirection;
foo:HDOP ?JasmintHDOP;
foo:Temperature ?JasminTemperature;
pos:alt ?JasminAltitude;
pos:long ?long;
pos:lat ?lat;

FILTER(?date >= "2012-02-07"^^xsd:date
&& ?date <= "2012-02-15"^^xsd:date)}}
```

Listing (6) **When did Jasmin go near the plantation on the 2012-02-15 and how close it was?**

```
SELECT DISTINCT ?date ?JasminTime ?Distance
{ GRAPH <urn:Jasmin> {?s
foo:GMTDate ?date;

foo:GMTTime ?JasminTime;
pos:long ?long;
pos:lat ?lat.
?s1 geof:nearby (5.612 117.8436 100 unit:Kilometer).
# Assume the plantation geoloction is (5.612 117.8436).
BIND (geof:distance(?s, ?s1, unit:Kilometer)
as ?Distance)
FILTER(?date = "2012-02-15"^^xsd:date}}
```

Listing (7) **What are the soil metrics near Jasmin between 2012-02-15-07?**

```
SELECT * {?s a sosa:Observation;
foo:Land_Use   "Logged␣Forest"^^xsd:string;
foo:Site   ?Site_Name;
foo:soil_Moisture  ?Moisture;
foo:soil_pH  ?pH;
foo:total_C  ?total_C  ;
foo:total_N  ?total_N.
SERVICE <db://FOO> {?Jasmin foo:gMTDate ?date.
FILTER (?date >= xsd:date("2012-02-07")
\&& ?date < xsd:date("2012-02-15"))}}
```

Listing (8) **What are the other elephants near the oil palm plantation?**

```
SELECT DISTINCT * {?s
foo:GMTDate ?date;
foo:GMTTime ?Time;
pos:long ?long;
pos:lat ?lat.
?s1 geof:nearby (5.612 117.8436 100 unit:Kilometer).
# Assume the plantation geoloction is (5.612 117.8436).
BIND (geof:distance(?s, ?s1, unit:Kilometer)
as ?Distance)
FILTER(?date = "2012-02-15"^^xsd:date)}
```

Listing 9. **Prefixes for CQs SPARQL queries**

```
Prefixes
foo: <https://w3id.org/xxx/xxx#>

time: <http://www.w3.org/2006/time#>
schema: <http://schema.org/>
xsd: <http://www.w3.org/2001/XMLSchema#>

dgfc: <http://schema.org/DGFC/elephant#>
pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>

geo: <http://www.opengis.net/ont/geosparql#>

geof: <http://www.opengis.net/def/function/geosparql/>
unit: <http://qudt.org/vocab/unit#>
```

Listing 10. **Reasoning (IF-Then-Rule) injected in FOO database**

```
prefix rule: <tag:stardog:api:rule:>
[] a rule:SPARQLRule ;
IF {
 ?s     pos:long ?long;
        pos:lat ?lat;
        <http://schema.org/DGFC/elephant#Jasmin>
        ?Jasmin.
 ?s1
        geof:nearby (5.612 117.8436
50 unit:Kilometer).
        BIND (geof:distance(?s, ?s1, unit:Kilometer)
        as ?Distance).
        BIND (?Distance <= 50 as ?Hazard).
        FILTER(?Hazard = TRUE)}
Then {
     ?s  pos:Hazard    ?Hazard}
```

5.5 Results

We conducted an experiment to assess the responses to four competency questions based on correctness, completeness, and speed. In the Stardog Studio (cloud.stardog.com) knowledge graph platform, we executed the queries 50 times each without reasoning, with a 3-second interval between queries, and recorded the response times. Subsequently, we repeated the experiment after enabling reasoning by activating the reasoning option and incorporating Semantic Web rule language (SWRL) into the triple store. The SWRL rule specifies that if the distance between an elephant and oil palm plantation is less than 50 km, it poses a hazard. Figure 16 illustrates the response times for each query, with average times of 54.7 ms, 87.29 ms, 259 ms, and 2080 ms for CQ 5, CQ 6, CQ 7, and CQ 8, respectively. Notably, the answers were accurate, with CQ 8 exhibiting the longest response time owing to the extensive information requested per query. Similarly, CQ 7 took longer to respond than CQ 5 and CQ 6 because of the need to connect disparate databases. Despite faster response times after reasoning, the correctness and accuracy of the query results remained consistent. We conducted the Shapiro-Wilk test for normality and the non-parametric Mann-Whitney U test to compare the two sets of independent responses. The Shapiro-Wilk test indicated that the 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, we opted for the Mann-Whitney U test, which revealed a significant difference between the query responses before and after enabling reasoning or inserting SWRL. Documentation for this evaluation is accessible at *https://github.com/xxxxxx/Knowledge-Graphs-Evaluations.git.*

Table 16. Statistics for the four competency questions 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	

Listing 11: 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 12: 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.
```

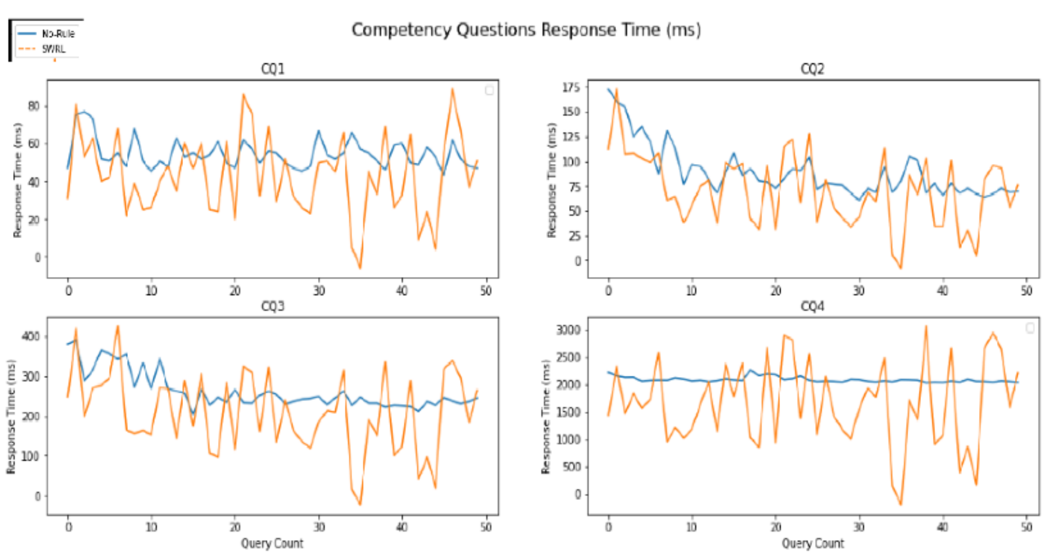


Fig. 16. Response time for the four competency questions

6 DISCUSSION

FOODS were generally well-received by domain experts, with high scores in confidence, usefulness, performance, time-saving, UI clarity, visualization, and data handling across participants with diverse educational backgrounds and occupations, mainly researchers and data specialists. Our knowledge graphs' ease of use and the need for technical support received more varied responses, indicating areas that few end-users might require prior indication training. The educational backgrounds and occupations of the participants suggest the knowledge graphs are relevant to academic and professional research, especially in fields requiring data integration and analysis. Our FOODS answered over 100 Competency Questions (CQs) and addressed three specific use cases. By promoting compatibility between different data types, the system facilitates more detailed analysis of wildlife data. Applying reasoning, particularly through Semantic Web Rule Language (SWRL) rules, optimized the inferred new knowledge from data and accelerated the query response times while maintaining accuracy. Through thorough statistical analyses of the results from the third use case, we proved that our system surpasses traditional data management methods, particularly in terms of speed and efficiency, thereby expediting the search and discovery processes. We evaluated FOO with the goal of improving decision-making by integrating various data sources. This evaluation involved the use of automated tools such as Oops! and insights from domain experts, thereby revealing minor structural and semantic issues in the ontology.

Challenges related to data sensitivity and scarcity were encountered, particularly concerning the ethical implications of GPS collar data for elephants. This emphasizes 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 our current framework relies on historical data, we recognize 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, such as 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 particularly useful 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. We 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, centered on defining domain-specific ontologies followed by the data population to generate FOODS, offers a flexible and replicable method across various domains. This approach is applied in healthcare [57], smart cities for urban planning [66], finance for market predictions [117], cultural heritage for connecting historical dots [28], and education for personalized learning solutions [22]. This methodology not only aids in structuring domain knowledge but also facilitates the extraction of actionable insights, demonstrating its broad applicability and potential to revolutionize knowledge representation and decision-making across diverse fields.

7 CONCLUSION

This study introduced a novel research data solution for forest observatories. It employed semantic web technologies to merge diverse wildlife research data that were previously separate. By creating the Forest Observatory Ontology (FOO) to construct knowledge graphs, a robust platform was established for detailed wildlife data analysis. The development of FOO was informed by qualitative analysis, which took into account interactions with various wildlife species and researchers in the forest. This ontology-driven method integrates data from various sources, enhancing advanced data analysis capabilities. Our proposed FOODS were evaluated using qualitative and quantitative methods to ensure their effectiveness and usability. Incorporating reasoning rules within FOO to detect potential threats, such as poaching, along with creating a user-friendly website, highlights this study's practical applications. Looking ahead, there is a commitment to expand research to include predictive modeling, which can further aid decision-making in wildlife conservation. These efforts signify promising avenues for future research, with the potential to transform conservation strategies and contribute to the sustainable management of biodiversity. Integration with Other Knowledge Graphs

REFERENCES

[1] Adnin, R., Haque, I., Afroz, S., Ishmam, A. M., Sarkar, S., Khan, M. K., Mimi, A., Chellappan, S., and Al Islam, A. A. (2023). Focusing on the unfocused: Corresponding perspectives on connectivity among small-scale non-profit organizations working for street children in bangladesh and their donors. *ACM Journal on Computing and Sustainable Societies*, 1(1):1–29.

[2] Ahmeti, A., Schakel, J.-K., David, R., and Revenko, A. (2023). Towards preserving biodiversity using nature first knowledge graph with crossovers. In *ISWC (Posters/Demos/Industry)*.

[3] 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.

[4] Allemang, D. (2020 - 2020). *Semantic web for the working ontologist : effective modeling for linked data, RDFS, and OWL*. ACM books ; 33. Association for Computing Machinery, New York, New York?, third edition edition.

[5] 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.

[6] 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.

[7] 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).

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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).

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] 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.

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

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

[36] 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.

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

[38] Fernández-López, M., Poveda-Villalón, M., Suárez-Figueroa, M. C., and Gómez-Pérez, A. (2019). Why are ontologies not reused across the same domain? *Journal of Web Semantics*, 57:100492.

[39] 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.

[40] 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.

[41] 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.

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

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

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

[45] 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.

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

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

[48] 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.
[49] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2).
[50] Grüninger, M. and Fox, M. S. (1995). Methodology for the design and evaluation of ontologies.
[51] Gutiérrez, C. and Sequeda, J. F. (2021). Knowledge graphs. *Communications of the ACM*, 64(3):96–104.
[52] 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.
[53] 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.
[54] 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*.
[55] 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).
[56] 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.
[57] 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.
[58] 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.
[59] 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.
[60] 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.
[61] 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.
[62] Keet, C. M. (2019). The african wildlife ontology tutorial ontologies: requirements, design, and content. *ArXiv*, abs/1905.09519.
[63] Keet, C. M. (2020). The african wildlife ontology tutorial ontologies. *Journal of Biomedical Semantics*, 11(1).
[64] Kejriwal, M. (2019). *Domain-specific knowledge graph construction*. Springer.
[65] 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.
[66] 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.
[67] Krötzsch, M. (2017). Ontologies for knowledge graphs? In *Description Logics*.
[68] 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.
[69] Lynn, M. S. and Jumail, A. (2021). The danau girang field centre: Field station profile. *ECOTROPICA*, 23(1/2):202103–202103.
[70] 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.
[71] Mansour, E., Srinivas, K., and Hose, K. (2022). Federated data science to break down silos [vision]. *ACM SIGMOD Record*, 50(4):16–22.
[72] McDaniel, M. and Storey, V. (2019a). Evaluating domain ontologies: Clarification, classification, and challenges. *ACM computing surveys*, 52(4):1–44.
[73] McDaniel, M. and Storey, V. C. (2019b). Evaluating domain ontologies: clarification, classification, and challenges. *ACM Computing Surveys (CSUR)*, 52(4):1–44.
[74] Michel, F., Djimenou, L., Zucker, C. F., and Montagnat, J. (2017). *xR2RML: Relational and non-relational databases to RDF mapping language*. PhD thesis, CNRS.
[75] Mireku Kwakye, M. (2019). Semantic data warehouse modelling for trajectories. *arXiv e-prints*, pages arXiv–1904.
[76] 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.

[77] Novo, O. and Francesco, M. D. (2020). Semantic interoperability in the iot: Extending the web of things architecture. *ACM Trans. Internet Things*, 1(1).
[78] Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology.
[79] 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.
[80] Page, R. D. (2019). Ozymandias: a biodiversity knowledge graph. *PeerJ*, 7:e6739.
[81] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.
[82] 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.
[83] 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.
[84] 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.
[85] 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.
[86] 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*.
[87] 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.
[88] Presutti, V., Blomqvist, E., Daga, E., and Gangemi, A. (2012). Pattern-based ontology design. *Ontology Engineering in a Networked World*, pages 35–64.
[89] 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*.
[90] 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.
[91] 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.
[92] 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.
[93] Ryen, V., Soylu, A., and Roman, D. (2022). Building semantic knowledge graphs from (semi-)structured data: A review. *Future Internet*, 14(5).
[94] 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.
[95] Sagar, S., Lefrançois, M., Rebaï, I., Khemaja, M., Garlatti, S., Feki, J., and Médini, L. (2018). Modeling smart sensors on top of sosa/ssn and wot td with the semantic smart sensor network (s3n) modular ontology. In *ISWC 2018: 17th Internal Semantic Web Conference*, pages 163–177.
[96] 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.
[97] Seydoux, N., Drira, K., Hernandez, N., and Monteil, T. (2016). Iot-o, a core-domain iot ontology to represent connected devices networks. In *European knowledge acquisition workshop*, pages 561–576. Springer.
[98] Singh, A. and Anand, P. (2013). State of art in ontology development tools. *International Journal*, 2(7).
[99] 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.
[100] 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*.

[101] Spivak, J. and Associate, T. (2013). NASA QUDT Handbook Ontology-based Specification of Quantities, Units, Dimensions and Types The 15th NASA-ESA Workshop on Product Data Exchange Colorado Springs, USA, 21-23 May 2013 Ralph Hodgson TopQuadrant CTO and NASA QUDT Ontologies Lead.

[102] 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.

[103] 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.

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

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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*.

[110] 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.

[111] 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].

[112] 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.

[113] 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.

[114] Wieder, P. and Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in Big Data*, 5.

[115] 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.

[116] 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.

[117] 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.

[118] Zou, X. (2020). A survey on application of knowledge graph. *Journal of Physics: Conference Series*, 1487(1):012016.

A APPENDIX: METHODOLOGY DETAILS

A.1 Ontology Requirements Specification Document (ORSD)

A.2 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.

A.3 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).

A.4 Implementation Language

The Web Ontology Language (OWL2) is used to implement FOO.

A.5 Intended End-Users

- Bioscientists.
- Wildlife Researchers.
- Computer Scientists.
- Data Scientists.

A.6 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.

A.7 Ontology Requirements

A.7.1 Non-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.

A.7.2 Functional Requirements.

- 106 curated Competency Questions (CQs),
- 10 Natural Language Statements (NLSs), see table 19.

	Open Data		Sensor Data	
Competency Questions (CQs)	Soil	Veg	GPS	Image
CQ1 Where do elephants forage?		*	*	
CQ2 What are the day-to-day movement patterns for elephant x?			*	
CQ3 What are the year-to-date (date range) movement patterns for elephant x?			*	
CQ4 What are the relationships between elephant x's movements and human/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?	*	*	*	
CQ8 What do elephants x, y, z's movements tell us?			*	
CQ9 How does elephant x use habitat site y?		*	*	
CQ10 What is the range of habitat sites used by elephants x, y, z?		*	*	
CQ11 Where was elephant x's location 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 x near 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 travelled between each stop (sleeping) by elephant x?			*	
CQ19 Which elephants met this month?			*	
CQ20 Which sites were revisited by elephant x this month?			*	
CQ21 What environment/habitat does elephant x prefer? Based on the prolonged time spent in a certain area.	*	*	*	
CQ22 Is there any significant change in elephant x's movement patterns between June and July 2015?			*	
CQ23 Has elephant x visited village y this year?			*	
CQ24 What is the movement range of elephant x during month y?			*	
CQ25 What is elephant x's activity (speed) during month y?			*	
CQ26 Are there any interactions between collared elephants during the flood season?	*		*	
CQ27 What is elephant x's tracking collar's battery activity?			*	
CQ28 What habitat does elephant x select this season?	*	*	*	
CQ29 What is the average elevation of elephant x during a specific time range?			*	
CQ30 Which elephant came near a logged site?		*	*	
CQ31 Which elephant came near a semi-logged site?		*	*	
CQ32 Which elephants crossed the river?			*	
CQ33 What is the canopy height for the distance travelled 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 travelled by elephant y in a specific period of time?			*	
CQ37 What is the altitude of the collared elephants?			*	
CQ38 What are the body/environment temperatures for collared elephants?			*	
CQ39 What is the behaviour 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 favourite foods in a particular area?		*	*	
CQ43 Do we need to create corridors along rivers/palm plantations, or is it not an obstacle for elephants to roam across the river?			*	
CQ44 What are the elephants collars fitted for almost two years?			*	
CQ45 What are the migration patterns for elephant x during the flood season?			*	
CQ46 What are the favourite 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?			*	
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?			*	
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?			*	*
CQ59 Is elephant x sick, injured, or dead?			*	*
CQ60 Which elephants are likely to conflict with humans?			*	
CQ61 What is the soil condition during a certain time of the year?	*			
CQ62 What are the types of soil available across the year? Dry, muddy, swamps	*			
CQ63 What are the soil characteristics with more or less nutrients, minerals?	*			
CQ64 What are the locations (soil type) that elephants prefer?	*			
CQ65 What are the mineral content (salt and others) in a particular location?	*			

Table 17. 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	Image
CQ66 Is there any metal in the soil in that area?	*			
CQ67 What are the chemical and agro-chemical concentrations in the soil of a certain area?	*			
CQ68 Does the soil in location x contain disease pathogens?	*			
CQ69 Which area needs pesticide spraying?	*			
CQ70 What is the soil moisture level in a specific location?	*			
CQ71 What is the presence of minerals in the soil?	*			
CQ72 Are there signs of heavy metals in the soil?	*			
CQ73 Where are the salt licks located?	*			
CQ74 What are the mineral and salt concentrations in the soil that indicate the presence of salt licks in a particular location?	*			
CQ75 What is the pH level of the soil?	*			
CQ76 What is the temperature reading from the soil sensor?	*			
CQ77 What is the soil moisture in a certain location?	*			
CQ78 Is the soil in this area healthy for animals?	*			
CQ79 Is the soil fertile in this area?	*			
CQ80 What is the moisture rate of the soil in this area (i.e., provide geo-location)?	*			
CQ81 Where to plant crops for elephants (i.e., soil moisture rates)?	*			
CQ82 Could planting in safer areas (healthy soil) influence animal movements?	*			
CQ83 Could we predict crop yield based on soil data?	*			
CQ84 What soil metrics help us predict flooding?	*			
CQ85 What are the metrics of healthy soil with less/no chemical pollution from oil palm plantations?	*			
CQ86 Why do elephants not like to walk on wet soil (movement prediction)?	*			
CQ87 What are the chemical levels in the soil?	*			
CQ88 What are the soil nutrient levels?	*			
CQ89 What is the effect of moisture on nutrients and oxygen levels?	*			
CQ90 What is the ideal soil moisture rate for an elephant to give birth?	*			
CQ91 What are the soil conditions in the areas that have elephant grass?	*			
CQ92 How to conserve suitable soils for the elephants to have food in the future (e.g., reduce using fertiliser)?	*			
CQ93 What soil moisture do elephants spend most time on?	*			
CQ94 Is it based on the plants grown in that soil?	*			
CQ95 How does this compare to urban areas? Or logged areas?		*		
CQ96 What do elephants eat?		*		
CQ97 Where are elephant grass (Napier), bark, palm shoots, young leaves, trunks, soft plants, and bananas?		*		
CQ98 Where do bamboo shoots grow?		*		
CQ99 Where could we find areas with the inner trunk of oil palms?		*		
CQ100 Where could we find areas with broad leaves?		*		
CQ101 Where could we find areas with vines?		*		
CQ102 How do vegetation and site habitat information help understand elephant behavior?		*		
CQ103 Do elephants drink lots of water?			*	
CQ104 Where do we find fruit farms in lower Kinabatangan?		*		
CQ105 What areas have fewer trees?		*		
CQ106 What plant species to conserve in the areas the elephants visit?		*	*	

Table 18. Table1-Continued

	Open Data		Sensor Data	
Natural Language Statements (NLSs)	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 19. Natural language statements and what data set can fulfil the task.

Fig. 17. Research visit and data collection at DGFC