Journal of Ecohumanism
2024
Volume: 3, No: 8, pp. 12375 – 12387
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.5845

Constructing Knowledge Graphs for New Species in the Greater Mekong Subregion

Yuttana Jaroenruen[1], Nattapong Kaewboonma[2], Wei-Ning Cheng[3]

Abstract

This study presents a methodology for developing a knowledge graph by integrating World Wildlife Fund (WWF) species discovery reports in the Greater Mekong Subregion. The research employs a four-phase approach: constructing a taxonomic data model, implementing a graph database, developing a web application architecture, and creating an interactive user interface. The knowledge graph integrates 2,783 species across 27 taxonomic groups, comprising 12,089 nodes and 30,303 connections. The system enables complex queries through SPARQL endpoints, revealing patterns in species distribution and conservation status from 1997 to 2022. Primary limitations include geographical constraints to the Greater Mekong Subregion and limited integration with external biodiversity databases. The system provides researchers with a graph search engine for exploring species data, supporting scientific research and conservation planning. This study presents a novel approach to organizing biodiversity data by combining taxonomic hierarchies with conservation metrics, establishing a framework for integrating species discovery data.

Keywords: *Knowledge Graph, Semantic Network, Species Discoveries, Taxonomy, Ontology, Mekong River.*

Introduction

Knowledge graphs can greatly benefit bioinformatics applications (Hamed, Rana, Goossens, Orozco-terWengel, & Perera, 2023) (Abu-Salih, 2021). A knowledge graph is a social network of resources such as ideas, concepts, and entities interconnected by semantic relations represented as a network of nodes linked by directed links (Khoo, Tan, Ng, & Chan, 2024). The information presented by nodes (i.e., concepts/entities) and edges (i.e., links) in the graphs, also known as node-link diagrams, can help to visualize computational biology ideas and enhance the understanding of domain knowledge for end-users. Ontologies, which are fundamental for representing domain knowledge, are typically used to construct these node-link diagrams. However, knowledge graphs are often created using existing datasets, and ontologies' importance must be addressed (Li, et al., 2024). We assert that a well-built ontology is crucial for establishing a knowledge graph. Firstly, ontologies may be organized based on taxonomies in some research (Salatino, et al., 2020); however, they also include rules specifying permissible relations between concepts/entities, which differ from simply categorizing different types of concepts (i.e., taxonomies). Furthermore, (Khoo, Tan, Ng, & Chan, 2024) emphasized that knowledge graphs should focus on supporting information-seeking and user behavior from an end-user perspective. A well-established ontology is necessary for machines to construct knowledge graphs automatically and for users to better understand domain knowledge by accessing and synthesizing information from visualized knowledge graphs.

Information should be provided by Knowledge Graph resources such as EOL (Encyclopedia of Life) (Sharafeldeen, Algergawy, & Konig-Ries, 2019), Wiki species, and Global Biodiversity Information Facility (GBIF) (García-Roselló, González-Dacosta, & Lobo, 2023). For example, what area is newly discovered, is it at risk of extinction, and to what extent? However, the above resources need to contain the information as an example. The Greater Mekong Subregion (GMS) is a biodiversity hotspot where new species are regularly discovered. Information about new species discoveries is often found in scientific journals and reports from organizations like the World Wildlife Fund (WWF) (Finlayson, 2018). We, therefore, close this gap by using a direct study of primary literature. Resources used in this project were found in WWF,

[1] Informatics Innovative Center of Excellence, School of Informatics, Walailak University, Email: jyuttana@wu.ac.th
[2] Faculty of Management Technology, Rajamangala University of Technology Srivijaya, Email: nattapong.k@rmutsv.ac.th, (Corresponding Author), Tel.: +66-8180-66110

[3] Graduate Institute of Library and Information Studies, National Taiwan Normal Univesity, Email: ivanka@ntnu.edu.tw

publications, journals, and authors who discovered or contributed to those new species in Greater Mekong Subregion documents. This paper presents a concept paper of valuable approaches that can be used to create a knowledge graph using information from WWF new species discoveries reports. Unlike other methods, the construction relies on ontology to help us produce data graphs.

Furthermore, we fill a gap in existing methods by discussing how to mix taxon relationships, and WWF categories may be inferred using graph properties. In addition, we show how end users can operate on this species discovery using a graph search engine via a Metaphactory platform (Eberhart, Haase, & Schell, 2023) and complete information on all nodes and relationship attributes. Finally, we provide some interesting statistics of new species discovered in the GMS that we create from data in the knowledge graph.

The Greater Mekong Subregion (GMS) is a natural economic region with a population of around 333.8 million people, connected by the Mekong River. It covers an area of 2.6 million square kilometers. It is home to over 20,000 plant species, 1,300 bird species, a thousand reptile and amphibian species, and over 500 mammal species (Berg, Lan, Da, & Tam, 2023). Since 1997, around 3,000 new species have been described in the GMS. The region is home to rare, vulnerable, and unique species, such as crested gibbons, forest pheasants, box turtles, the Irrawaddy dolphin, and the secretive Saola (World Wide Fund for Nature, 2017) (World Wide Fund for Nature, 2021). However, the region is at a crossroads due to the unprecedented dangers posed by a rising human footprint, consumerism, and unsustainable economic development activities. In 2020, new species like the Popa langur, cavefish, succulent bamboo, and iguana were introduced globally, highlighting the importance of conservation efforts in the GMS (Arthan, et al., 2023).

Related Works

Research related to knowledge graphs for new species discovery includes specifying a conceptualization of the new species. A semantic network represents a network of objects, events, situations, or concepts and illustrates the relationship between them. In the framework of the domain top-level ontology BioTop, Stefan et al. (Stefan, Holger, & Martin, 2008) presented an ontological approach to biological taxa, and BioTopLite, a simplified version of BioTop, contains 55 classes and 37 properties with streamlined relations for better usability (Schulz, Boeker, & Martinez-Costa, 2017). It is founded on the premise that every biological creature, population, or biological stuff has some taxon characteristic intrinsic to it. This unique concept is wholly integrated into the Open Biological Ontology standard and aligned with the BFO (Masseroli, 2018), a significant top-level ontology. The technique depicts taxon attributes as a simple is_a hierarchy, making it simple and scalable to ingest portions of existing taxonomy databases like the NCBI taxonomy. This data may be automatically converted into an OWL (Web Ontology Language) subtype hierarchy and automatically connected to the BioTop node TaxonQualty.

Furthermore, (Chansanam, Kwiecien, Buranarach, & Tuamsuk, 2021) developed a thesaurus for ethnic groups in the Mekong River Basin, a collection of controlled vocabularies in Thai and English, and a digital platform that allows for semantic search and connected open data. The research followed the Research and Development Method, which included four steps: (1) analysis, synthesis, and knowledge organization; (2) thesaurus building; (3) the development of a digital thesaurus platform; and (4) the assessment of the digital thesaurus platform. The Tematres online program was used to create the digital thesaurus. The constructed thesaurus of ethnic groups in the Mekong River Basin has 4273 fundamental terms in two languages (Thai and English). There are 2596 words with hierarchical relationships, broader and narrower terms, and 6858 with associative or connected terms.

The research on African Wildlife Ontology (AWO) instructional ontologies, introduced by (Keet, 2020), AWO serves as an educational framework, meeting 22 critical requirements for ontology development and surpassing previous models in alignment and competency evaluation, ultimately improving ontology engineering education. For another animal, Bird Ontology employs Protégé 5.5.0-beta-9 to classify birds into Paleognathae (flightless) and Neognathae (flying), prioritizing data integration and classification over instructional use, with online visualization through ontograf and OWL viz (Nandhinidevi, Saraswathi, Thangamani, & Ganthimathi, 2021).

More examples of research on taxonomy were created by (Maria, James, Naila, Alyona, & Jacob, 2021) using a systematic examination of the wildlife laws of eight countries in the Global South: Angola, Brazil, Cambodia, Costa Rica, Indonesia, Kenya, Mexico, and Vietnam. These were picked to represent various countries with distinct legal systems and outstanding biodiversity. They created a taxonomy with a 4-level hierarchical structure using iterative sorting and reduction. Offenses, including the harvest, transportation, use of wildlife, forgeries, and obstruction of justice, were all included in Level 1 of the taxonomy. After that, each category was subdivided into mutually exclusive subcategories. The Taxonomy of Wildlife Offenses yielded 511 offense categories organized into a four-level hierarchy. The 511 offenses are organized into a hierarchical taxonomy that researchers and practitioners may use for legal analysis. This is relevant in light of opposing efforts to strengthen, deregulate, and modify wildlife legislation, particularly concerning concerns about zoonotic dangers and widespread biodiversity loss. Page (Page, 2016) adopted JSON-LD for biodiversity data, created reconciliation services to match entities to IDs, and used a combination of document and graph databases to store and query the data to develop a helpful biodiversity knowledge graph. To launch this project, they can build wrappers for each significant biodiversity data provider and a central cache as both a document store and a straightforward graph database. Applications that use the central cache to address particular issues, such as enhancing existing data, should highlight the effectiveness of this strategy.

Lastly, the study of Ozymandias, a biodiversity knowledge graph that includes information on Australian animal species and their names, taxonomic publications, their authors, and their interrelationships, such as publication, citation, and authorship. Over 9 million triples comprise the knowledge graph, implemented as a triple store. This triple store's online interface allows users to browse the data from numerous angles, such as taxonomy, publications, and authors. This interface was created as part of the GBIF Ebbe Nielsen Challenge 2018 (Roderic, 2019).

Materials and Methods

This section presents the proposed system's technical features and information content. First, we introduce the arrangement of the species data in the Greater Mekong Subregion and the data sources we consider. Second, we describe the ontology classes used to manage and consolidate all the data. Finally, we describe the structure of the graph database, the query language used, and the individual attributes of related data.

Materials

The new species ontology in the GMS comprises the following biological and bioinformatics resources. WWF project reports from 1997 to 2022. The document is a report on the species in the GMS, organized into 12 groups according to the WWF report's title: First Contact (World Wide Fund for Nature, 2007), Close Encounters (World Wide Fund for Nature, 2008), New Blood (World Wide Fund for Nature, 2010), Extra-terrestrial (World Wide Fund for Nature, 2012), Mysterious Mekong (World Wide Fund for Nature, 2014), Magical Mekong (World Wide Fund for Nature, 2021), Species Oddity (World Wide Fund for Nature, 2016), Stranger Species (World Wide Fund for Nature, 2016), New Species on The Block (World Wide Fund for Nature, 2018), New Species Discoveries (World Wide Fund for Nature, 2021) and Primate of Greater Mekong (World Wide Fund for Nature, 2021), New Species Discoveries in the Greater Mekong 2021 & 2022 (World Wide Fund for Nature, 2023) in order of publication. The Encyclopedia of Life (EOL) is a complimentary online encyclopedia that documents all the 1.9 million living species known to science. It is built from existing authorized databases that experts and non-experts have edited worldwide (Parr, et al., 2016). Wikispecies is a wiki-based online initiative backed by the Wikimedia Foundation. The project aims to create a comprehensive open content library of all species, focusing on scientists rather than the public. The GNU Free Documentation License and the Creative Commons Attribution-Share Alike 3.0 License apply to Wikispecies (Karipidis & Prentzas, 2020).

However, we have encountered some resource data extraction issues, which are divided into two problems. The first is that WWF's resources are printed media and do not present biological data. Secondly, knowledge graph data sources that we have used, such as EOL and Wiki species, are presented with biological data but need more information on species conservation, discovery, and discoverers, as well as spatial and ecological data.

To solve such problems, we require a comprehensive look at several data sources, including:

Unstructured public literature sources, such as WWF's reports to identify specific categories of Greater Mekong Subregion and Binomial or Scientific name

Structured external sources, such as EOL databases, to understand the expression profile of the species

We might explain how our text analytics and semantic enrichment can assist with the new GMS species occurrences.

To effectively construct new GMS species instances, we planned to manage the complexity and mess of existing data and external references. All the previously referenced sources are available in print and online. Most of the data files received are available in the documents as text. We used species names as identifiers in a document analysis method and stored data in CSV formats. After entering the species' preliminary information described in the WWF's reports, we used the GREL script for semantically enriching (Angelis & Kotis, 2021) to organize data coherence and suitable linkages between entities. It then links other referenced resources not included in the WWF's documents. As for the next step, we have used that information to develop an in-house taxonomy named GMKS taxonomy (It is different from GMS, which means Greater Mekong Subregion.) by the Tematres web application (https://vocabularyserver.com/web). The aim is to ensure that this glossary is required to identify all the terms an indexer can use to index documents. We have analyzed the terminology to index bibliographic records, called the Descriptors, which are then used as the basis for the taxonomy and become an integral part of the controlled vocabulary. The Descriptors will become the starting point and conduct the indexer to select related and narrow words.

The GMKS taxonomy contains 2,783 descriptor terms across four levels: Level 1 (6 terms), Level 2 (74 terms), Level 3 (22 terms), and Level 4 (2,681 terms). The taxonomy uses three types of semantic relationships (equivalence, hierarchy, and BT/NT/USE/UF) and includes 75 Classes, 23 Relationships, and 14 Attributes. The taxonomy can be displayed alphabetically and in classified sequences based on concept relationships.

Methods

Our research process has four steps: creating a data model, building a data graph, designing a web application architecture, and creating a web user interface. The components of each step are shown in the figure below.

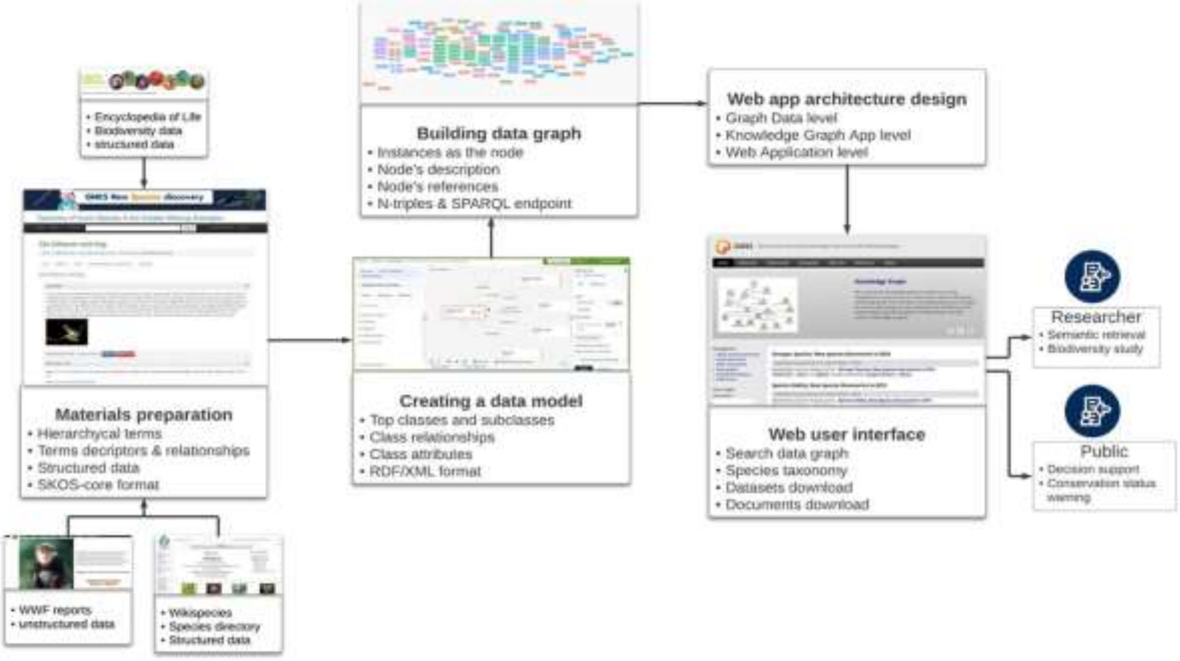


Figure 1. GMKS Knowledge Graphs Development Framework

Creating A Data Model

Developing knowledge graph data modeling is the process we undertake to describe species names as a connected graph of nodes and relationships of the top domains we are interested in studying. The broad arrangement of the knowledge graph is based on the EOL Identifiers and biodiversity data from Wikispecies. The core entities are species names, species groups classified by WWF, publications, journals, and authors who discovered or contributed to those species in GMS. Almost all entities and references in our datasets obtained from the GMKS taxonomy are well-defined. As a result, providing a conceptual description of the GMKS database is relatively simple, as seen in Figure 3. Our model followed a general concept: The GMKS knowledge graph would map any biological entity into a qualifying node. The relationship between two or more data entities is mapped as a relationship using the species name as an identifier.

The Species and Info classes are based on the GMKS taxonomy's Descriptor. Species represent nodes in a taxonomic tree as instances of rdfs:species. The RDFS structure uses the species_info_of relationship to reflect parent-child classifications. Our data model integrates the taxon concept and WWF species discovery documents, featuring two properties: subclass_of (identifying species from WWF documentation) and category_of (using common names in species groups like bats, insects, and orchids).

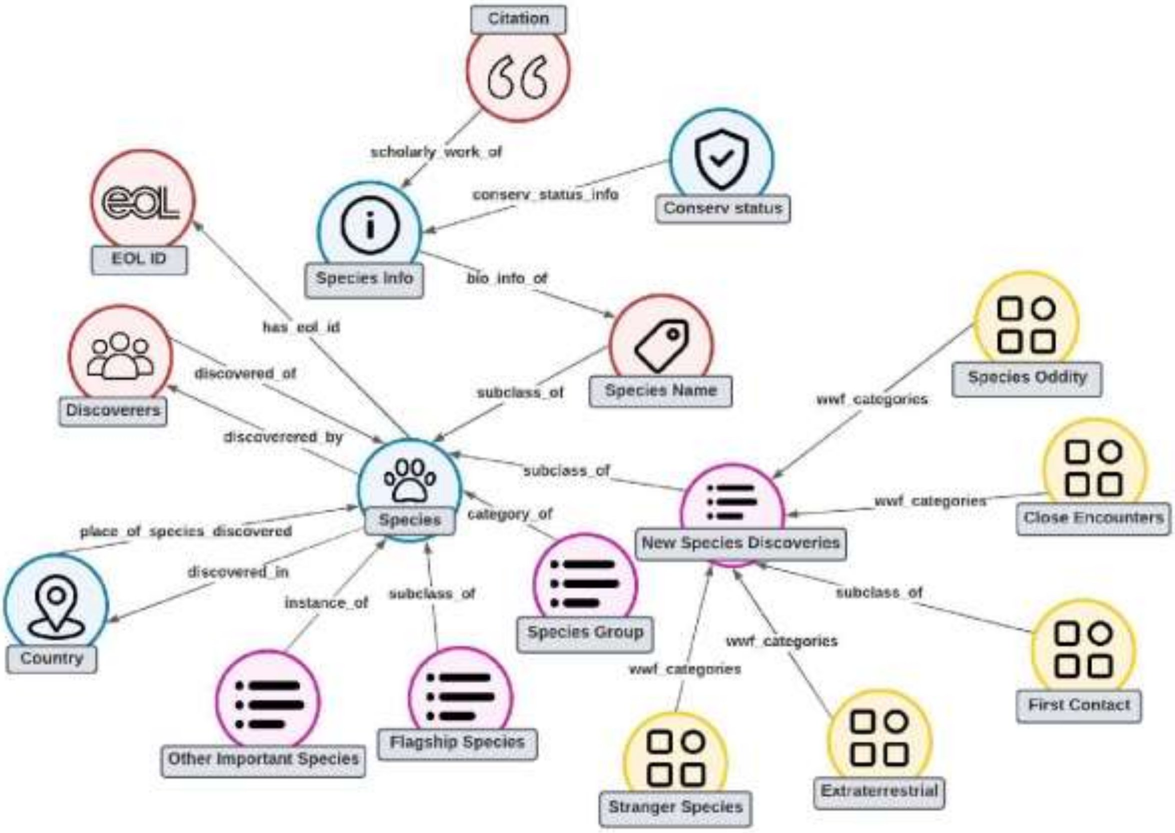


Figure 2. The Knowledge Graph Model Used in GMKS

Entities are categorized into labeled classes, such as Species Info and Species Name, with a relationship between them established using GMKS taxonomy. We also created classes like Species_Name and Conserve_status based on WWF classifications. In our current dataset, these subclasses are linked to the Species_info class. Our model includes five main classes, 75 subclasses, 19 relations, and 14 attributes, comprising 12,089 nodes and 30,303 connections.

Species_Info class: Represents information about species within the GM Subregion, including subclasses for Species_Name, Citation, Discoverers, EOL_ID, and Full-Text Link

Species class: This category defines units of biological classification through three WWF categories: Flagship Species (for conservation support), Other Important Species, and New Species Discoveries (focused on new species found from 1997 to 2022).

Species_Group class: Introduces common names for various species (e.g., fish, bats, orchids) to make biological information accessible to non-scientists.

Country and Conserv_status class: This class includes a Country class for geographical data on species discoveries in the Greater Mekong Subregion and a Conserv Status class providing conservation details from the International Union for Conservation of Nature. New relationships derived from WWF documentation include conserv_status and wwf_category, linking nodes involved in species conservation and biodiversity study.

Building Data Graph

The process begins with structuring the ontology, comprising five main classes, 75 subclasses, 19 relations, and 14 attributes. This structure organizes essential information about species, discoverers, locations, and conservation status.

Node creation follows, where each node contains three key components: Annotations (species details and references), Types (species classification), and Relationships (node connections). These nodes form the foundation of the knowledge graph, establishing connections through fundamental relations like "is_a Plant," "discovered_by," "discovered_in," and "conserv_status."

The technical implementation involves installing metaphactory 4.5.0 on a Linux server and importing data as N-triples using named graphs. The system provides access through REST service, SPARQL endpoint, and a web interface (http://lod.walaiautolib.com/), followed by testing and verifying data relationships and queries to ensure system reliability.

Web Application Architecture Design

The architecture of the web application is modular. The application has a full-stack architecture, as shown in Figure 3. We also use metaphacts technology (Xu, Curé, & Calvez, 2021) on our servers to ensure responsiveness and efficiency. There are three levels of tools used. From the bottom to the top, the graph data level, the knowledge graph application level, and the web application level are handled by metaphactory. Other cross-functional and cross-communication methods and data formats are also covered.

The Graph Data level consists of the SPARQL 1.1 graph database (Gupta & Malik, 2022), other data sources, and data processing services. It allows users to run queries to graph instances stored within it remotely. The data is retrieved for operators to manage at level 2, the Knowledge Graph Application level. This tier includes Knowledge Graph Management, End-user Oriented Interaction, and Knowledge Graph Application modules.

The last level is the Web Application level. An online user interface is primarily assembled with HTML, CSS, JSON, and JavaScript frameworks to make the creation of responsive web front-ends simple and quick. We also used the metaphactory platform to manage dynamic content manipulation and asynchronous information transfers to the data graph. Ultimately, we developed a web application for retrieving and displaying new species discoveries in the Greater Mekong Subregion. The data and display are updated regularly by the system. Visitors can try browsing by visiting our search page at http://lod.walaiautolib.com/resource/resource and logging in as a guest.

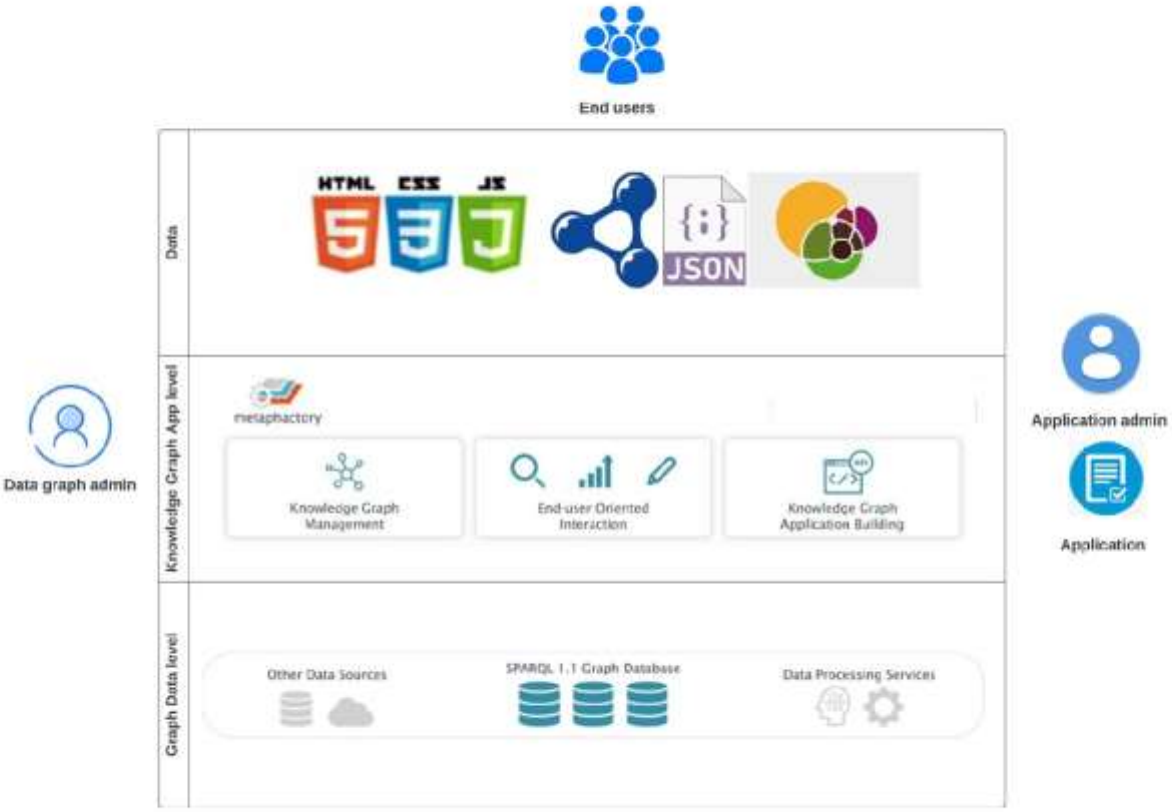

End users
Data
HTML
CSS
JS
JSON
Knowledge Graph App level
metaphactory
Knowledge Graph Management
End-user Oriented Interaction
Knowledge Graph Application Building
Graph Data level
Other Data Sources
SPARQL 1.1 Graph Database
Data Processing Services
Data graph admin
Application admin
Application

Figure 3. GMKS Architectural Stack

Web User Interface

The GMKS knowledge graph web interface offers extensive species information across seven sections. The Home page welcomes visitors, while the Search data graph section features a powerful search engine for accessing node and relationship attribute data. The Visual Taxonomy section illustrates relationships of newly discovered species in the Greater Mekong Subregion alongside the GMKS species taxonomy on the Tematres vocabulary server. Users can find primary resources in the Documents section and download the GMKS taxonomy and knowledge graph in various formats from the Datasets section. The About Us section details projects and staff, making the interface a comprehensive resource for understanding regional species.

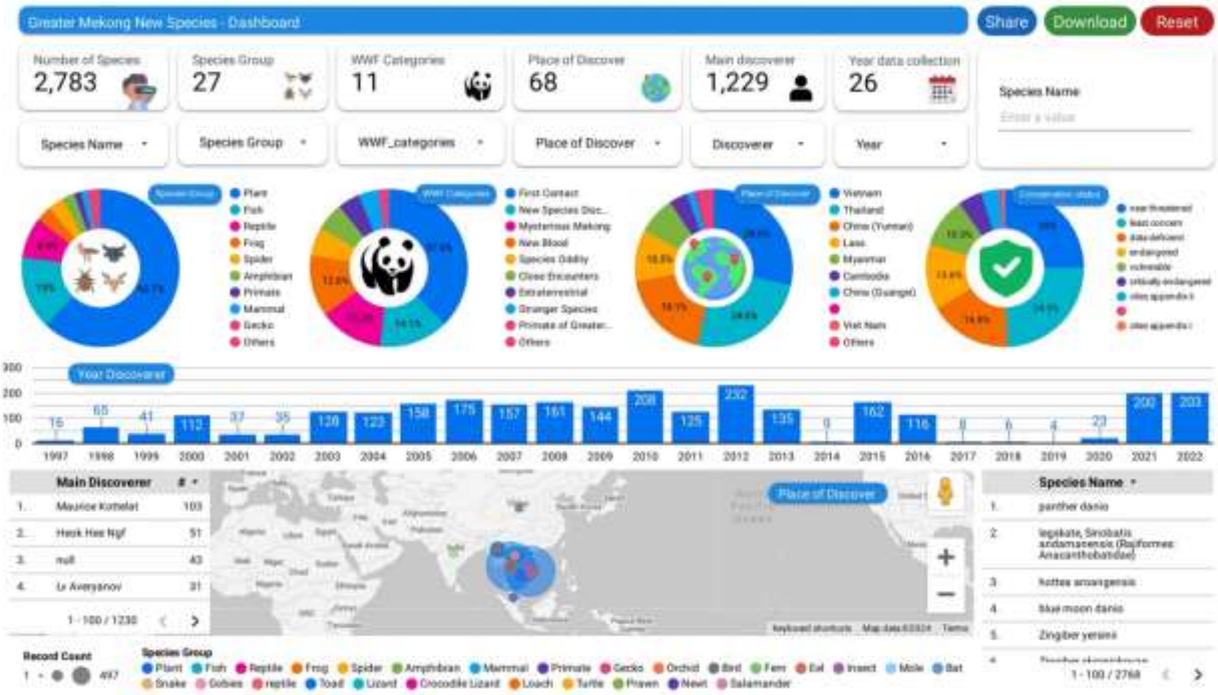
System Validation

The validation process for the GMKS knowledge graph integrated standard ontology utilized NetworkX for network analysis (Chen, et al., 2022). It began with RDF/XML ontology file validation, followed by analyzing class hierarchies, property definitions, and instance relationships, revealing a scale-free network typical of biological knowledge systems. The validation confirmed the graph's structural integrity, showing no inheritance issues or circular dependencies. Network metrics indicated effective knowledge organization through clustering and centrality measures. This dual validation method ensured the ontology's technical accuracy and the knowledge network's practical utility for biodiversity research in the Greater Mekong Subregion.

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.5845

Results

We can perform some exploratory analyses by utilizing the underlying knowledge graph. For example, users can use the knowledge graph for various experimental investigations, defined and outlined below. They can also use the SPARQL queries to generate the results listed in the supporting information section.



Number of Species	Species Group	WWF Categories	Place of Discover	Main discoverer	Year data collection
2,783	27	11	68	1,229	26

Species Name ▾ Species Group ▾ WWF_categories ▾ Place of Discover ▾ Discoverer ▾ Year ▾

Species Name
Enter a value

	Main Discoverer	# ▾
1.	Maurice Kottelat	103
2.	Heok Hee Ng\f	51
3.	null	43
4.	Lv Averyanov	31

1 - 100 / 1230 ‹ ›

	Species Name ▾
1.	panther danio
2.	legskate, Sinobatis andamanensis (Rajiformes: Anacanthobatidae)
3.	hottea arnangensis
4.	blue moon danio
5.	Zingiber yersinii
6.	[illegible]

1 - 100 / 2768 ‹ ›

Figure 4. GMKS Summary Statistics

Summary Statistics

Figure 4 shows a screenshot of the data set general statistics. The GMKS knowledge graph encompasses comprehensive data about species discoveries in the Greater Mekong Subregion, documenting 2,783 species across 27 species groups and 11 WWF categories. The data spans 68 geographic areas and covers 26 years of collection (1997-2022), with contributions from 1,229 principal discoverers. This extensive dataset provides a robust foundation for analyzing biodiversity patterns and trends in the region.

Knowledge Graph Statistics

Overall, our knowledge graph shows healthy characteristics in terms of connectivity and structure, but it could benefit from a more balanced relationship distribution. The NetworkX as validation tools suggest it's functional but has room for refinement in how information is interconnected. The following key metrics characterize the interconnectedness of the network, as shown in Table 1.

Table 1. Knowledge Graph Statistics

Metric	Value
Total Nodes	12,089
Total Edges	30,303
Average Node Degree	5.01
Maximum Node Degree	4,711

Metric	Value
Network Density	0.0004
Number of Connected Components	3
Largest Component Size	12,079
Average Clustering Coefficient	0.1487
Minimum Degree	1
25th Percentile	1
Median Degree	1
75th Percentile	6

Figure 5 shows the view of a knowledge graph for a species named Acranthera burmanica. We see the associate species of the WWF species group (plant), the people who discovered this species (Yun-hong Tan), its status of conservation (near threatened), and the country (Myanmar) where the researcher found it. We also located lists of Encyclopedia of Life Identifiers (EOL ID) associated with the species names and the link to scholarly work (identifiers) that made those names public.

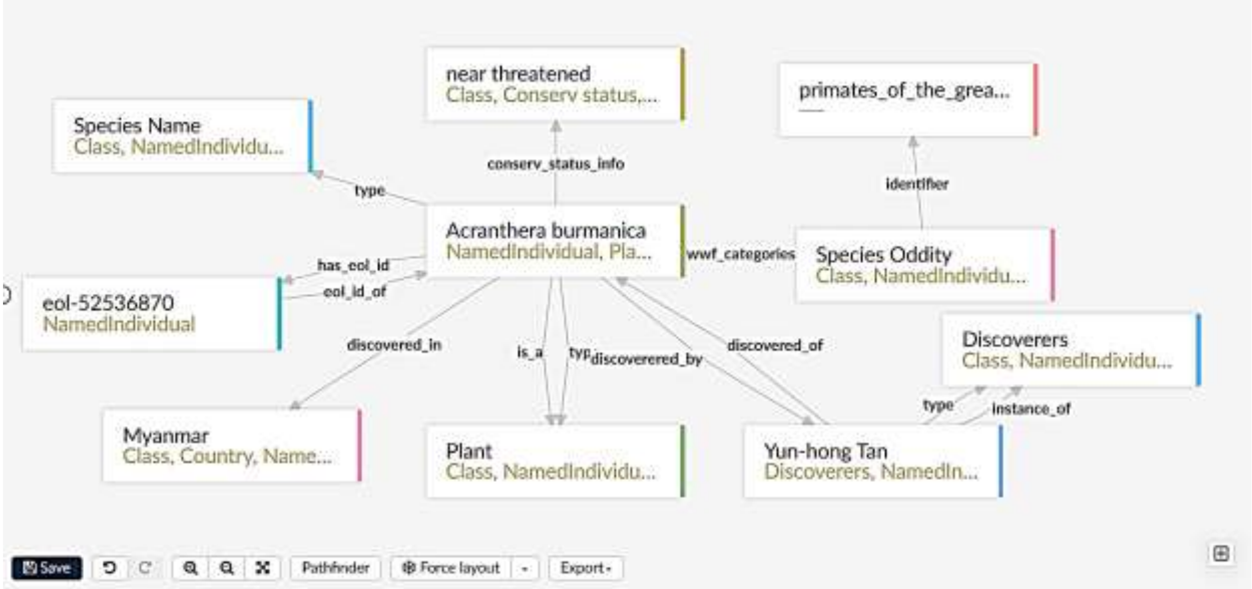


Figure 5. An Example of Species Acranthera Burmanica Search Results from Metaphact

Taxonomic Structure and Species Distribution

The knowledge graph incorporates a comprehensive taxonomic framework comprising 75 classes, including 70 subclasses and eight classes with subclasses. The structure is enriched with 29 total relations (object properties), and 14 attribute properties, of which 13 have defined domains and 14 have specified ranges. This robust framework enables detailed classification and relationship mapping across the dataset.

Analysis of species discoveries in the Greater Mekong Subregion from 1997 to 2022 reveals distinct distribution patterns across various dimensions. Taxonomically, plants constitute the majority at 62.1% of all discoveries, followed by fish at 15% and reptiles at 8.4%, with other species groups accounting for the remaining 14.5%. Geographically, Vietnam leads with 28.6% of discoveries, followed by China's Yunnan region (24.6%), Myanmar (19.1%), and Thailand (10.5%), with other locations comprising 17.2% of discoveries.

From a conservation perspective, the species demonstrate varying levels of vulnerability. A quarter (25%) of the species are classified as Near Threatened, while 24.5% are Least Concern. Notably, 16.8% are categorized as Vulnerable, 13.6% as Endangered, and 10.3% as Data Deficient, with the remaining 9.8%

classified as Critically Endangered or falling into other conservation categories. These distributions highlight the region's rich biodiversity and its pressing conservation challenges.

Data Accessibility and Integration

The knowledge graph integrates multiple data sources, comprising 1,229 principal discoverers, 27 species groups, 11 WWF categories, 68 discovery locations, and data collected over 26 years. This integration facilitates efficient querying and analysis of species data, thereby supporting research and conservation efforts. The graph effectively organizes complex biological data while ensuring accessibility for both scientific and general audiences. As a result, it serves as a valuable tool for biodiversity research in the Greater Mekong Subregion.

Discussion

The GMKS knowledge graph system demonstrates several unique characteristics compared to existing biodiversity knowledge organization systems. Unlike BioTop/BioTopLite (Stefan, Holger, & Martin, 2008), which focuses primarily on biological taxonomy through a simple is_a hierarchy, GMKS integrates taxonomic classification and conservation data, providing a more extended view of species information in the Greater Mekong region.

While African Wildlife Ontology emphasizes educational aspects and wildlife data from Africa (Keet, 2020), GMKS takes a broader approach by combining species discovery tracking, conservation status, and geographical context specific to the Greater Mekong Subregion. This integration allows for a more detailed analysis of regional biodiversity patterns and conservation needs.

Compared to the Ethnic Groups Thesaurus (Chansanam, Kwiecien, Buranarach, & Tuamsuk, 2021), which specializes in bilingual cultural data management, GMKS focuses on scientific species data while maintaining accessibility through its user-friendly interface and multiple visualization options. Similarly, whereas Ozymandias concentrates on publication and citation networks for Australian biodiversity (Roderic, 2019), GMKS provides a more holistic view by combining WWF reports, EOL data, and Wiki species information.

The key advantage of GMKS lies in its comprehensive integration of multiple data types: taxonomic classification, conservation status, discovery information, and geographical context. This integration, with its accessible interface and visualization capabilities, makes it particularly valuable for biodiversity research and conservation efforts in the Greater Mekong Subregion. Additionally, GMKS's ability to track new species discoveries provides an individual temporal dimension not recognized in other systems.

Table 2. Comparison of Knowledge Graph Systems for Biodiversity and Taxonomic Data

System	Data Sources	Methodology	Output/Interface	Unique Strengths
BioTop and BioTopLite	- NCBI taxonomy - Existing databases	- BFO-based modeling - is_a hierarchy	- Taxonomic trees - OWL hierarchy	- Strong formal ontology - Biological focus
African Wildlife Ontology	- Wildlife data - Educational content	- Educational framework - Protégé development	- Educational interface - Teaching tools	- Educational focus - African wildlife expertise
Ethnic Groups Thesaurus	- Cultural data - Bilingual content	- 4-step R&D process - Thesaurus building	- Digital platform - Semantic search	- Bilingual support - Cultural context
Ozymandias	- Publications - Citation data	- Triple store - Network analysis	- Web interface - Multiple views	- Publication focus - Australian biodiversity

GMKS	- WWF reports (primary) - EOL data - Wiki species - Conservation data	- 4-step process - Combined taxonomy-ontology - UI development	- Web interface - Visual taxonomy - Multiple visualizations	- Conservation integration - Discovery tracking - Geographic context - User accessibility

Conclusions

We offer our GMKS, which focuses on constructing taxonomies and knowledge graphs through two main phases of development. The first phase involved creating the GMKS taxonomy based on a systematic analysis of WWF's primary resources. This phase included defining terms as concepts and their relationships and developing a taxonomy web application. The second phase focused on developing the knowledge graph. This involved determining the data model, establishing connections between species information, designing a three-tier web architecture, and constructing a user interface.

The GMKS successfully integrated multiple data sources, covering 2,783 species across 27 species groups and 11 WWF categories. Its four-level taxonomy structure features partitive and instance relationships, making it suitable for ontology and linked data migration. Furthermore, the GMKS knowledge graph explicitly classifies specific types of new species in the Greater Mekong Subregion, including flagship species, recent discoveries, and other significant species. It also provides valuable information on species and conservation status classes for further research

References

Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. Journal of Network and Computer Applications, 185(1 July), pp. 1-21. doi:https://doi.org/10.1016/j.jnca.2021.103076

Ahmeti, A., Schakel, J.-K., David, R., & Revenko, A. (2023). Towards Preserving Biodiversity using Nature FIRST Knowledge Graph with Crossovers. CEUR Workshop Proceedings, 3632.

Angelis, S., & Kotis, K. (2021). Generating and Exploiting Semantically Enriched, Integrated, Linked and Open Museum Data. Communications in Computer and Information Science, Volume 1355 CCIS, 367 - 379. doi:10.1007/978-3-030-71903-6_34

Arthan, W., Ohrnberger, D., Sungkaew, S., Phosi, S., Teerawatananon, A., & Janloy, A. (2023). A new species and a new record of Bambusa (Poaceae: Bambusoideae) from Thailand. Kew Bulletin, 78(4), 597 - 606. doi:10.1007/s12225-023-10137-5

Berg, H., Lan, T. H., Da, C. T., & Tam, N. T. (2023). Stakeholders assessment of status and trends of ecosystem services in the Mekong Delta for improved management of multifunctional wetlands. Journal of Environmental Management, 338(July 2023). doi:10.1016/j.jenvman.2023.117807

Chansanam, W., Kwiecien, K., Buranarach, M., & Tuamsuk, K. (2021). A Digital Thesaurus of Ethnic Groups in the Mekong River Basin. Informatics, 8(3). doi:https://doi.org/10.3390/informatics8030050

Chen, Y., Liu, J., Xian, M., Wang, H., Zhang, Y., Y, Z., . . . Zhang, L. (2022). Construction of network security domain knowledge graph for network attack detection. 6th International Conference on Electronic Information Technology and Computer Engineering, EITCE 2022. October 2022, pp. 1171 - 1178. Virtual, Online: Association for Computing Machinery. doi:10.1145/3573428.3573638

David, N., & Casey, S. (2020). Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal, 18, 1414-1428.

Eberhart, A., Haase, P., & Schell, W. (2023). metaphactory for Massive Graphs. 14th Annual ACM/SPEC International Conference on Performance Engineering, ICPE 2023 (pp. 215 - 220). Coimbra: Association for Computing Machinery, Inc. doi:10.1145/3578245.3585330

Finlayson, M. (2018). World wide fund for nature (WWF). Springer Netherlands. doi:10.1007/978-90-481-9659-3_139

García-Roselló, E., González-Dacosta, J., & Lobo, J. M. (2023). The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently. Biological Conservation, 283(July 2023). doi:10.1016/j.biocon.2023.110118

Gupta, R., & Malik, S. K. (2022). An analysis of SPARQL usage for information retrieval in heterogeneous domains through various tools. Journal of Discrete Mathematical Sciences and Cryptography, 25(4), 1031 - 1040. doi:10.1080/09720529.2022.2072428

Hamed, N., Rana, O., Goossens, B., Orozco-terWengel, P., & Perera, C. (2023). FOO: An Upper-Level Ontology for the Forest Observatory. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, (pp. pp. 154 - 158). doi:10.1007/978-3-031-43458-7_29

Karipidis, N., & Prentzas, J. (2020). Main features and types of educational use of wiki technology. In Handbook of Research on Modern Educational Technologies, Applications, and Management (pp. 213 - 228). IGI Global. doi:10.4018/978-1-7998-3476-2.ch013

Keet, C. M. (2020). The African wildlife ontology tutorial ontologies. Journal of Biomedical Semantics, 11(4), 1-11.

Khoo, C. S., Tan, E. A., Ng, S.-G., & Chan, C.-F. (2024). Knowledge Graph Visualization Interface for Digital Heritage Collections: Design Issues and Recommendations. Information Technology and Libraries, 43(1), 1-26. doi:https://doi.org/10.5860/ital.v43i1.16719

Li, X., Luo, R., Liu, K., Li, F., Wang, C., & Wang, Q. (2024). A New Approach for Ontology Generation from Relational Data Design Patterns. 8th International Conference on Data Mining and Big Data, DMBD 2023. 2018 CCIS, pp. 157 - 172. Sanya: Advanced Institute of Big Data, Beijing, China. doi:10.1007/978-981-97-0844-4_12

Maria, P., James, W., Naila, B., Alyona, R., & Jacob, P. (2021). Building a global taxonomy of wildlife offenses. Conservation Biology, 35(6), 1903-1012.

Masseroli, M. (2018). Integrative bioinformatics. In Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics (Vols. 1-3, pp. 1092 - 1098). Elsevier. doi:10.1016/B978-0-12-809633-8.20388-9

Nandhinidevi, S., Saraswathi, K., Thangamani, M., & Ganthimathi, M. (2021). Design and development of bird ontology using protégé. doi:10.1016/j.matpr.2021.01.596

Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal, 18, 1414-1428. doi:https://doi.org/10.1016/j.csbj.2020.05.017

Page, R. (2016). Towards a biodiversity knowledge graph. Research Ideas and Outcomes, 2(e8767). doi:https://doi.org/10.3897/rio.2.e8767

Parr, C. S., Schulz, K. S., Hammock, J., Wilson, N., Leary, P., Rice, J., & Corrigan, R. J. (2016). TraitBank: Practical semantics for organism attribute data. Semantic Web, 7(6), 577 - 588. doi:10.3233/SW-150190

Roderic, D. M. (2019). Ozymandias: a biodiversity knowledge. PeerJ-Life and Environment, 7((e6739)), 1-25.

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., & Motta, E. (2020). The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intelligence, 2(3), 379 - 416. doi:10.1162/dint_a_00055

Schulz, S., Boeker, M., & Martinez-Costa, C. (2017). The BioTop Family of Upper Level Ontological Resources for Biomedicine. Studies in Health Technology and Informatics, 235, 441 - 445. doi:10.3233/978-1-61499-753-5-441

Sharafeldeen, D., Algergawy, A., & Konig-Ries, B. (2019). Towards Knowledge Graph Construction using Semantic Data Mining. 21st International Conference on Information Integration and Web-Based Applications and Services, iiWAS 2019. Munich: Association for Computing Machinery. doi:10.1145/3366030.3366035

Stefan, S., Holger, S., & Martin, B. (2008). The ontology of biological taxa. Bioinformatics, 24(13), pp. i313–i321.

World Wide Fund for Nature. (2007). First Contact in the Greater Mekong: New Species Discoveries. Hanoi: WWF Greater Mekong Programme.

World Wide Fund for Nature. (2008). Greater Mekong: Close Encounters, New Species Discoveries 2008. Chanthabouly: WWF Greater Mekong Programme. Retrieved from https://wwfasia.awsassets.panda.org/downloads/greater_mekong_new_species_08_compressed.pdf

World Wide Fund for Nature. (2010). New Blood: Greater Mekong New Species Discoveries 2009. [n.p.]: WWF Greater Mekong Programme.

World Wide Fund for Nature. (2012). Extra Terrestrial: Extraordinary New Species Discoveries in 2011 from the Greater Mekong. Hanoi: WWF-World Wide Fund for Nature.

World Wide Fund for Nature. (2014). Mysterious Mekong: New Species Discoveries 2012-2013. Bangkok: WWF Greater Mekong Programme.

World Wide Fund for Nature. (2016). Species Oddity: New Species Discoveries in 2015. [n.p.]: WWF-World Wide Fund for Nature.

World Wide Fund for Nature. (2016). Stranger Species: New Species Discoveries in 2016. [n.p.]: WWF-World Wide Fund for Nature.

World Wide Fund for Nature. (2017). Greater Mekong. Retrieved from New Species Discoveries In The Greater Mekong: https://www.worldwildlife.org/places/greater-mekong/

World Wide Fund for Nature. (2018). New Species on the Block in 2017. [n.p.]: WWF-World Wide Fund for Nature.

World Wide Fund for Nature. (2021). Magical Mekong: New Species Discoveries in the Greater Mekong Region in 2014. Bangkok: WWF Greater Mekong Programme.

World Wide Fund for Nature. (2021). New Species Discoveries in the Greater Mekong 2020. WWF-World Wide Fund for Nature.

World Wide Fund for Nature. (2021). Primates of the Greater Mekong: Status, Threats and Conservation Efforts. [n.p.]: WWF-World Wide Fund for Nature.

World Wide Fund for Nature. (2023). New Species Discoveries in the Greater Mekong 2021 & 2022. Bangkok: WWF-Greater Mekong.

Xu, W., Curé, O., & Calvez, P. (2021). Knowledge graph management on the edge. Advances in Database Technology - 24th International Conference on Extending Database Technology, EDBT 2021. 2021-March, pp. 229 - 240. Nicosia: OpenProceedings.org. doi:10.5441/002/edbt.2021.21.