Preprint of article for 12th International Conference on Information and Communication Technologies in Agriculture, Food & Environment (HAICTA 2026)

proceedings

MDPI

Proceeding Paper

Connecting the Dots in Marine Biodiversity Using Ontology-Based Knowledge Graphs †

Yannis Marketakis [1,2] *, Eleni Tsouloucha [1], Athina Kritsotaki [1], Yannis Tzitzikas [1,2], Loida S. Corpus [3], Nicolas Baily [3], Josephine Rius-Barile [3], Vanessa Camille Agbay [3] and Noriel Porcincula [3]

[1] Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece; marketak@ics.forth.gr (Y.M.); tsoulouha@ics.forth.gr (E.T.); athinak@ics.forth.gr (A.K.); tzitzik@ics.forth.gr (Y.T.)
[2] Computer Science Department, University of Crete, Heraklion, Greece
[3] Quantitative Aquatics Inc., Los Banos, Philippines; l.corpus@q-quatics.org (L.C.); n.bailly@q-quatics.org (N.B.); j.barile@q-quatics.org (J.B.); v.agbay@q-quatics.org (V.A.); j.porcincula@q-quatics.org (J.P.)
* Correspondence: marketak@ics.forth.gr (Y.M.)
† Presented at the 12th International Conference on Information and Communication Technologies in Agriculture, Food & Environment (HAICTA|2026), Kos Island, Greece, 30 September – 3 October, 2026.

Abstract

Biodiversity data are essential for understanding ecosystems, species biology, and environmental change. FishBase is one of the most widely used information systems for fish species, providing rich data on taxonomy, morphology, ecology, and distribution. However, its relational structure limits interoperability with open and linked scientific data. This paper presents SemantyFish, a semantic knowledge graph derived from FishBase, enabling integration, reuse, complex queries, and cross-domain exploration across biodiversity and fisheries-oriented information systems, while aligning with FAIR data principles. A key contribution is an ontology-driven evolution framework that keeps the knowledge graph synchronized with the evolving FishBase database through schema mappings between the original relational schema and a unified ontology. To support access and reuse, SemantyFish is exposed through a RESTful API that allows humans and machines to query, browse, and retrieve resources. Overall, the proposed approach enhances the interoperability and scientific value of large-scale biodiversity databases.

Keywords: FishBase; knowledge graphs; biodiversity information systems; marine biodiversity; semantic data integration; FAIR data principles

1. Introduction

Biodiversity data are fundamental for understanding ecosystem functioning, species biology, and the impacts of environmental change. As global efforts intensify to monitor biodiversity and support sustainable resource management, the availability of reliable, high-quality, and interoperable data resources has become increasingly important. Over the last decades, several large biodiversity information systems have emerged. Among them, FishBase is one of the most widely used resources for marine and freshwater fishes, providing curated information on taxonomy, life history, ecology, distribution, and human uses. It aggregates expert-validated data for tens of thousands of species and is continuously updated as new knowledge becomes available. FishBase is also widely used in fisheries research, marine resource monitoring, and sustainability-oriented data

Academic Editor: Firstname Lastname

Published: date

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

workflows, making its interoperability particularly important for ICT applications in the broader agriculture, food, and environment. Despite its scientific value and global adoption, it is primarily implemented as a relational database designed for human browsing. This structure constrains its integration with other biodiversity and environmental data infrastructures, such as Global Biodiversity Information Facility (GBIF) [1], the Ocean Biodiversity Information System (OBIS) [2], and the Global Record of Stocks and Fisheries (GRSF) [3], and limits its use in machine-driven workflows supporting biodiversity monitoring, fisheries management, sustainability assessment and large-scale ecological analyses.

Semantic Web technologies [4] provide a foundation for addressing these limitations by representing data in machine-interpretable form using ontologies and standardized vocabularies. This enables the construction of knowledge graphs that support semantic interoperability, data linking, and complex queries across heterogeneous sources. Transforming FishBase into a semantic knowledge graph allows its biological content to be connected with complementary datasets, such as habitat descriptions, environmental observations, or food and nutrition information, while aligning the resource with the FAIR data principles [5].

This paper presents SemantyFish [6], an ontology-based framework for converting FishBase into a knowledge graph that supports both construction and long-term evolution. A top-level ontology for the marine domain provides the unifying conceptual foundation for modeling species, traits, habitats, and related information. Schema mappings between the relational sources and the ontology enable systematic construction and regeneration of the knowledge graph as the underlying database evolves. To facilitate access and reuse, the resulting knowledge graph is exposed through a RESTful API for both human and machine clients.

The main contributions of this work are: (a) an ontology-driven approach for transforming a large biodiversity database into a semantic knowledge graph, (b) a schema-mapping framework that ensures semantic consistency and interoperability, (c) an evolution mechanism that keeps the knowledge graph synchronized with a continuously changing source, and (d) an API layer that supports standardized access and integration with external biodiversity, fisheries, and environmental information systems.

2. Background and Related Work

Large-scale biodiversity information systems play a central role in documenting species, their distributions, and their biological and ecological characteristics. Major global infrastructures such as GBIF [1] and OBIS [2] aggregate and disseminate occurrence data from multiple providers, supporting research and conservation. Alongside these, FishBase [7] is a widely used resource for marine and freshwater fishes, offering curated information on taxonomy, ecology, distribution, and human uses. However, as a relational database designed primarily for human-oriented access, it has limited interoperability with other biodiversity and environmental data infrastructures.

Semantic Web technologies and ontologies enable data to be expressed in a machine-interpretable and semantically explicit form, supporting interoperability and integration across heterogeneous sources. In life sciences, initiatives such as the OBO foundry [8], and domain ontologies for taxonomy, anatomy and ecology have demonstrated the value of formal semantics for data integration. In the biodiversity domain, several efforts have explored publishing datasets as Linked Open Data [9], including semantic layers for GBIF and OBIS, as well as the use of Wikidata as a collaboratively curated biodiversity knowledge graph [10].

Knowledge graphs enable rich querying and cross-domain exploration by linking species data with habitats, environmental conditions, and scientific literature. However,

https://doi.org/10.3390/xxxxx

existing approaches typically provide static or loosely coupled semantic views of underlying databases, offering limited support for maintaining consistency as data evolve. In contrast, this work introduces an ontology-based knowledge graph designed for continuous synchronization with a dynamic source through schema-mapping-driven transformation, ensuring semantic interoperability and long-term sustainability in line with the FAIR data principles.

3. Methodology

This section outlines the methodological steps for constructing the knowledge graph, focusing on the conceptual approach, the schema mappings and evolution framework. Figure 1 shows the main steps and the evolution framework.

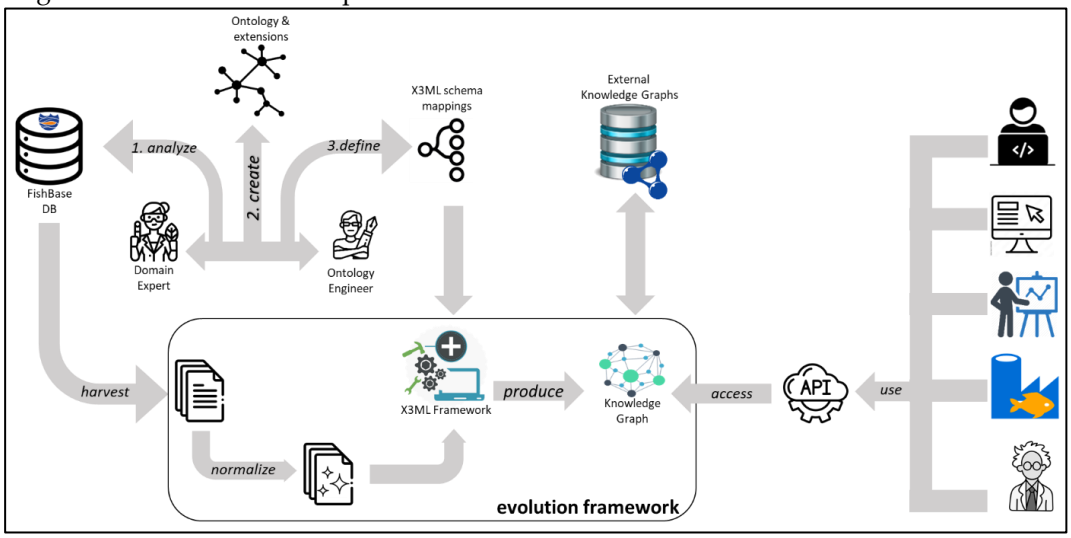


Figure 1. Methodology for constructing and maintaining the SemantyFish Knowledge Graph

3.1. Conceptual Approach

The conceptual design of the knowledge graph was driven by a systematic analysis of the FishBase relational schema to uncover the underlying domain semantics rather than its table structures. This analysis identified core conceptual entities such as taxonomic units, life-history traits, habitat descriptors, geographic distributions, and bibliographic references, together with their relationships. These elements form an implicit conceptual model in the original database, which is not explicitly represented in a way that supports interoperability or reuse beyond the FishBase environment.

To make this conceptual model explicit and reusable, an ontology-driven approach was adopted. MarineTLO [11], a top-level ontology for the marine domain, was selected as the semantic backbone of the knowledge graph, as it offers a coherent framework for representing organisms, habitats, ecological characteristics, observations, and provenance information in a unified and machine-interpretable form. Moreover, MarineTLO complements biodiversity exchange standards by providing a formal semantic layer for integration and reasoning.

FishBase-specific concepts and attributes were then aligned to this ontology through controlled extensions that capture domain-specific notions such as growth parameters, stock details, and distributional descriptors. This process resulted in a unified conceptual model that was used as the foundation for the definition of the schema mappings and the transformation workflow that is described below.

3.2. Schema Mappings

The alignment between the FishBase relational schema and the ontology-based conceptual model was realized through explicit schema mappings that define how source

data are transformed into instances of ontology classes and properties. Instead of relying on ad-hoc scripts or procedural transformations, the mappings are specified in a declarative manner that captures the intended semantics of the data independently of any particular implementation.

For this purpose, the X3ML framework [12] was used. X3ML provides a mapping language for defining schema mappings between heterogeneous data sources and ontologies in a structured and human-readable manner. It enables ontology experts and domain specialists to collaboratively specify how source schema entities correspond to ontology classes and properties, ensuring both technical and domain knowledge are preserved during the transformation. In addition, X3ML incorporates a transformation engine that applies the defined mappings to the source data to generate the knowledge graph. By separating mapping definition from the transformation logic, the framework supports systematic and repeatable generation of the knowledge graph, which is essential for the evolution framework described next.

3.3. Evolution Framework

FishBase is a continuously evolving database, with new species, traits, references, and corrections being regularly introduced by domain experts. In such a setting, a one-off semantic transformation would quickly become outdated, undermining the reliability and usefulness of the resulting knowledge graph. To address this issue, we implemented an evolution framework that treats the knowledge graph as a derived representation that can be systematically regenerated from the authoritative relational source. The regeneration process can be executed periodically or triggered by updates in the source database, enabling near real-time synchronization depending on operational requirements. Moreover, the framework supports partial reconstruction by selectively re-transforming selected FishBase tables. The corresponding resources are organized into separate named graphs, enabling efficient incremental updates without regenerating the entire knowledge graph.

The framework is based on the schema mappings defined in X3ML, which serve as a stable semantic contract between the FishBase schema and the extended MarineTLO ontology. Since the structure of the FishBase relational database is highly stable and changes only rarely, these mappings can be reapplied whenever the data are updated to generate a new version of the knowledge graph without requiring manual re-engineering. By decoupling the conceptual model from the transformation process, and anchoring the construction of the knowledge graph in declarative mappings, the framework supports both content and structural evolution. Updates to either the source schema or the ontology can be accommodated by revising the schema mappings, and re-executing the framework, enabling controlled, repeatable, and transparent synchronization between the two representations.

4. Knowledge Graph Implementation and Access

The ontology-based mappings and transformation workflow described above were applied to the FishBase database to generate a semantic knowledge graph that represents various biodiversity resources in a unified, machine-interpretable form. The resulting knowledge graph integrates heterogeneous FishBase tables into a coherent semantic structure, enabling consistent identification of entities and explicit representation of their relationships. Table 1 provides some statistics about the size and the contents of the knowledge graph so far[1], as well as the overall time that is required for constructing it.

[1] The knowledge graph is continuously expanded as additional tables and entities are progressively modelled and transformed. Therefore, the reported figures represent a snapshot at the time of writing the paper (July 2026)

https://doi.org/10.3390/xxxxx

Figure 2 shows marine species in the Mediterranean Sea that occur in brackish habitats deeper than 50 m, are commercially important and have at least one Greek common name. Results include the scientific name, Greek common name, depth, areas, and danger indicator. In the relational schema, answering this requires multiple joins, whereas the knowledge graph enables a declarative query over a unified conceptual model.

Table 1. Size and content statistics of the Knowledge Graph

Type	Size	Type	Size
RDF triples	20.7 million	**Synonym instances**	244,412
Distinct resources	5.7 million	**Stock instances**	36,629
Species instances	36,175	**Country instances**	307
Genera instances	11,838	**FAO areas instances**	26
Family instances	665	**Ecosystem instances**	36,363
Common name instances	386,456	**Species dimensions**	143,382
Ontology classes used	61	**Ontology properties used**	65
Transformation time	~ 22 minutes	**KG construction time**	~ 3 minutes

SPARQL | HTML5 table

species_scientific_name	common_name_in_greek	depth	med_area	dangerous_species_indicator
"Huso huso"	"Μουρούνα"	"180"	"Mediterranean Sea"	"harmless"
"Huso huso"	"Στουριόνι"	"180"	"Mediterranean Sea"	"harmless"
"Huso huso"	"Ακιπήσιος"	"180"	"Mediterranean Sea"	"harmless"
"Dasyatis pastinaca"	"Αγριόψαρο"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"
"Dasyatis pastinaca"	"Μούντριζα"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"
"Dasyatis pastinaca"	"Βατοτρυγόνα"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"
"Dasyatis pastinaca"	"Μούτροβα"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"
"Dasyatis pastinaca"	"Τρυγόνα"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"
"Dasyatis pastinaca"	"Παπλωματάς"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"
"Dasyatis pastinaca"	"Μούτρουβα"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"
"Dasyatis pastinaca"	"Σφυράτσα"	"200"	"Northeast Atlantic and Mediterranean Sea"	"venomous"

Figure 2. Results of a complex (SPARQL) query that combines multiple entities

In addition to internal integration, the knowledge graph supports interoperability with external biodiversity infrastructures through explicit linking of its entities to publicly available knowledge bases. Species and taxonomic entities are linked, where possible, to corresponding identifiers in external sources, such GRSF. These links enable the enrichment of FishBase entities with complementary information (e.g., stock and fisheries details) and allow seamless navigation across distributed biodiversity knowledge graphs.

To make the knowledge graph accessible and reusable, a RESTful API was implemented as an access layer on top of the knowledge graph. The API provides endpoints for searching, browsing, and retrieving resources such as species, taxonomy, synonyms, ecosystems, and stocks, supporting both human users and external systems. It abstracts the underlying semantic technologies and enables interaction through a standard web mechanism, returning results in JSON format.

This access layer supports queries that combine multiple criteria across entities, enabling use cases that are difficult to support with the original relational system. External applications can integrate the API to combine FishBase data with other biodiversity resources, facilitating interoperability. The SemantyFish API is accessible at https://demos.isl.ics.forth.gr/semantyfish-api/, while Figure 3 shows indicative screenshots.

https://doi.org/10.3390/xxxxx

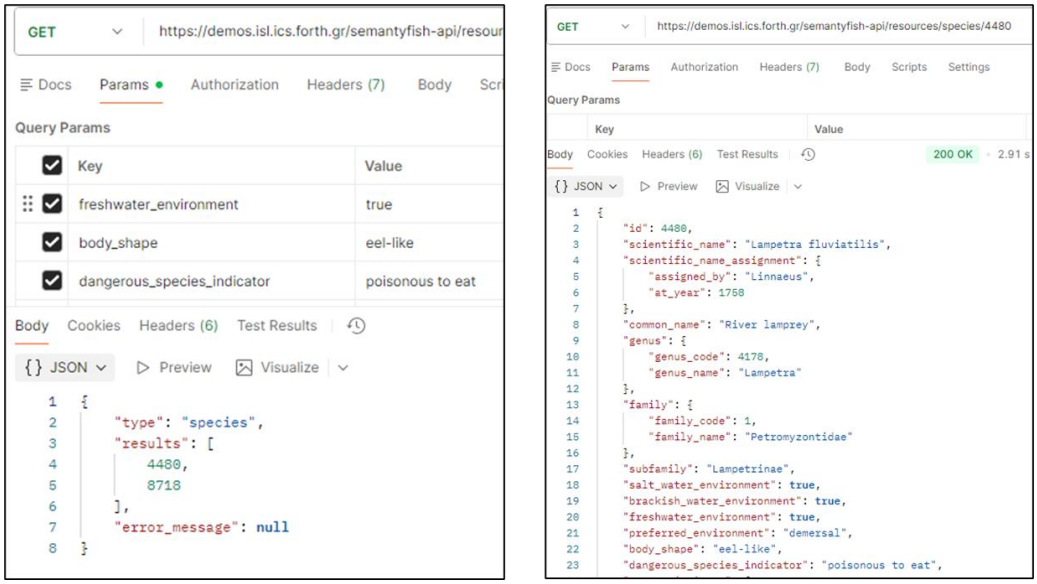


Figure 3. Screenshot of the SemantyFish API (left: search for species using specific criteria, right: show details of a particular fish species[2]).

5. Conclusion

This paper presented SemantyFish, a semantic knowledge graph derived from FishBase using ontology-driven modeling and Semantic Web technologies. By relying on a marine top-level ontology and declarative schema mappings, SemantyFish provides a machine-interpretable, interoperable, and FAIR-compliant representation that supports advanced querying and cross-domain integration. A central contribution is the evolution framework, which enables systematic reconstruction and synchronization of the knowledge graph as FishBase evolves, ensuring long-term consistency and sustainability. Through the SemantyFish API, external applications can access and integrate FishBase data in a standardized way, combining them with other biodiversity, fisheries, and environmental resources.

Regarding limitations, the current version covers only part of the FishBase schema, while additional entities and relationships are progressively incorporated. Although the conceptual model, ontology alignments, and schema mappings have been reviewed by domain and ontology experts, a formal semantic quality assessment is still needed. Future work will focus on extending coverage, increasing interlinking with external knowledge bases, evaluating semantic quality through expert validation, consistency checks, and enrichment through scientific literature and additional biodiversity sources.

Author Contributions: Conceptualization, Y.M., Y.T. and N.B.; methodology, Y.M.; software, Y.M.; validation, L.C., N.B., J.B., V.A and J.P.; formal analysis, E.T., A.K., N.B. and J.B.; investigation, Y.M.; resources, N.B. and J.B.; data curation, J.B., V.A. and J.P.; writing—original draft preparation, Y.M.; writing—review and editing, L.C., N.B. and Y.T.; visualization, Y.M.; supervision, Y.T. and L.C.; project administration, Y.M.; funding acquisition, Y.M. and N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded under OSCARS (Horizon Europe, grant agreement No. 101129751) 1st Call for Open Science Projects and Services.

Institutional Review Board Statement: Not applicable

[2] https://demos.isl.ics.forth.gr/semantyfish-api/resources/species/4480

Informed Consent Statement: Not applicable

Data Availability Statement: The data are publicly available through the API.

Conflicts of Interest: The authors declare no conflicts of interest

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
FAIR	Findable Accessible Interoperable Reusable
KG	Knowledge Graph
GBIF	Global Biodiversity Information Facility
GRSF	Global Record of Stocks and Fisheries
OBIS	Ocean Biodiversity Information System
OBO	Open Biological and Biomedical Ontologies

References

1. Lane, M.A. and Edwards, J.L., 2007. The global biodiversity information facility (GBIF). Systematics Association special volume, 73, p.1.
2. Grassle, J.F., 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. Oceanography, 13(3), pp.5-7.
3. Tzitzikas, Y., Marketakis, Y., Minadakis, N., Mountantonakis, M., Candela, L., Mangiacrapa, F., Pagano, P., Perciante, C., Castelli, D., Taconet, M. and Gentile, A., 2017, September. Towards a Global Record of Stocks and Fisheries. In HAICTA (pp. 328-340).
4. Antoniou, G. and Van Harmelen, F., 2004. A semantic web primer. MIT press.
5. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), pp.1-9.
6. SemantyFish: Advancing Visibility, Interoperability and Exploitability of FishBase. Available online: https://semanty-fish.github.io/ (accessed on 8 June 2026).
7. Froese, R. and D. Pauly. Editors. 2025. FishBase. World Wide Web electronic publication. www.fishbase.org, version (04/2025).
8. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J. and Obi Consortium, 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology, 25(11), pp.1251-1255.
9. Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L. and Desmet, P., 2014. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PloS one, 9(8), p.e102623.
10. Waagmeester, A., Schriml, L. and Su, A., 2019. Wikidata as a linked-data hub for Biodiversity data. Biodiversity Information Science and Standards, 3, p.e35206.
11. Tzitzikas, Y., Allocca, C., Bekiari, C., Marketakis, Y., Fafalios, P., Doerr, M., Minadakis, N., Patkos, T. and Candela, L., 2016. Unifying heterogeneous and distributed information about marine species through the top-level ontology MarineTLO. Program, 50(1), pp.16-40.
12. Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., Flouris, G. and Doerr, M., 2017. X3ML mapping framework for information integration in cultural heritage and beyond. International Journal on Digital Libraries, 18(4), pp.301-319.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.