Preprint of Y. Marketakis, Y. Tzitzikas, A. Gentile, A. Ellenbroek, M. Taconet, Building a Global Aquatic Resources Knowledge Base for Fisheries. Accepted for publication in proceedings of the 11th International Conference on ICT in Agriculture, Food & Environment Conference (HAICTA 2024), October 17-20, 2024, Samos, Greece

Building a Global Aquatic Resources Knowledge Base for Fisheries

Yannis Marketakis [1,2], Yannis Tzitzikas [1,2], Aureliano Gentile[3], Anton Ellenbroek[3] and Marc Taconet[3]

[1] *Institute of Computer Science, FORTH-ICS, Heraklion, Greece*
[2] *Computer Science Department, University of Crete, Heraklion, Greece*
[3] *Food and Agriculture Organization of the United Nations, Rome, Italy*

Abstract

Fisheries management is a complex task aiming to ensure the long-term sustainability of fish populations and the ecosystems they depend on. To achieve those goals, it is essential that the fisheries are described with precise and non-ambiguous information. Different agencies are reporting fisheries data by relying on several vocabularies or thesauri. Just indicatively for the description of aquatic species there are different official and widely used data sources that can be used. As a result, there are different identifiers or names for describing the same resource. In this paper, we describe the construction of a global aquatic resources knowledge base, that is the result of the integration of different data sources using semantic web technologies. By focusing on aquatic species, we show that the information provided by different data sources is complementary, and we provide a unified way for accessing them. We finally describe how the same process was adopted for other information domains as well.

Keywords

Fishery, Semantic Data Integration, Knowledge Base, Species, Taxonomy, Water Area, Fishing Gear

1. Introduction

Fishing is one of the most significant drivers of declines in ocean wildlife populations. Although catching fishes is not inherently bad for the ocean, except for when the catches are higher than the fish stocks can replenish, something called overfishing. The number of overfished fish stocks globally has tripled in half a century. Just indicatively, according to [1], one-third of global fish stocks were overfished in 2017. Fisheries management [2] plays a crucial role in preventing overfishing by regulating the harvesting and utilization of fish stocks to ensure their sustainable exploitation while preserving the marine ecosystem. It involves a range of activities aimed at maintaining the balance between the extraction of fishery resources and the conservation of aquatic ecosystems. One of the key components towards efficient fisheries management is the so-called stock assessment. This involves monitoring fish populations to determine their abundance, distribution and health. It is therefore of crucial importance, that the description of relevant information (e.g. species) is accurate and complete. There are many different ways for referring to a species; the majority of people use common names. These are not used from the scientific community, because species usually have many different common names in various languages. Instead, they are using the Linnean binomial nomenclature, or the so-called scientific name of a species; It consist of two parts, the genus name first and then the specific epithet (*Mullus barbatus*). In addition, there are alpha-numeric identifiers that can be used for that purpose. Such identifiers are provided from different data sources or registries and are widely used, particularly

Proceedings of HAICTA 2024, October 17–20, 2024, Samos, Greece
EMAIL: marketak@ics.forth.gr (A.1); tzitzik@ics.forth.gr (A.2); aureliano.gentile@fao.org (A.3); anton.ellenbroek@fao.org (A.4); marc.taconet@fao.org (A.5)
ORCID: 0000-0002-0417-2526 (A.1); 0000-0001-8847-2130 (A.2); 0000-0002-6542-132x (A.3); 0000-0002-9262-3103 (A.4); 0000−0002−3103−6204 (A.5)

for data exchange and interoperable mechanisms. Some indicative ones are FAO 3-Alpha Species Code[2] (e.g. *MUT*), APHIA ID from WoRMS[3] (e.g. *126985*), FishBase ID[4] (e.g. *790*), TSN code from ITIS[5] (e.g. *169419*), etc. The problem is that there is not any common guideline adopted from existing fisheries management authorities and other institutions spread around the world. So practically all of them are being used nowadays and there are cases in which can become rather cumbersome to analyze fishery reports produced by different authorities.
To alleviate this, in this paper we have implemented a process that relies on semantic web [3] by collecting information from different data sources, and constructing a single knowledge base with key species taxonomic information. By using an appropriate ontology, as the conceptual model we managed to semantically integrate information coming from different data sources and describe them in a homogeneous manner. This process managed to interconnect the identifiers of the same species and exactly because of this, to support the provision of complementary information. This is something that would not be possible at all without semantically integrating them.

2. Semantic Data Integration

Semantic data integration involves the harmonization of heterogeneous data sources by understanding the underlying semantics, relationships, and meanings within the data. It goes beyond syntactic matching to interpret the semantics of data elements, resolving semantic heterogeneity that arises from differences in terminologies and concepts across sources. This process employs ontologies [4], vocabularies and definition of schema mappings to establish common semantic interpretations across disparate datasets.

The key element is the definition of a proper ontology, able to capture clearly the semantics of the knowledge base of aquatic species. In the literature there are several ontologies for the marine domain, such as MarineTLO [5], SWEET [6], BioTop [7], Uberon [8], etc. In this work, we have adopted MarineTLO, since it contains all the necessary information we wanted to model and there was no need for extending it with new concepts, attributes or relations.

3. Marine Species Data Sources

For the construction of the knowledge base, we have used the following well-known and actively used data sources:

- FAO ASFIS List of Species for Fishery Statistics Purposes[2] provides a code list of marine species with several identifiers, such as 3-alpha code, taxonomic code and ISSCAAP code[6]. 3-alpha codes are made of three characters that uniquely identify the species, complemented with the scientific names, taxonomic details, and common name in English, French, Spanish, Russian, Arabic and Chinese when available. The taxonomic codes and ISSCAAP codes are other codes that are used for classificatory purposes.
- WoRMS is an authoritative database that provides a comprehensive and up-to-date inventory of all known species globally. The main identifier in WoRMS is AphiaID which is a numeric code. Moreover, it contains taxonomic information, synonyms, distribution maps, and bibliographic references for each species listed in the database.
- FishBase is a global biodiversity information system on fishes that provides detailed information about species regarding taxonomy, morphology, ecology, distribution, behavior, and fisheries-related data. FishBase is maintained and continuously updated by an international consortium of scientists, with support from various organizations.
- ITIS is an authoritative database that provides taxonomic information including the Taxonomic Serial Number (TSN), scientific names and taxonomic hierarchies. The database

[2] FAO ASFIS List of Species for Fishery Statistics Purposes (https://www.fao.org/fishery/en/collection/asfis/en)
[3] World Register of Marine Species (https://www.marinespecies.org/)
[4] FishBase (https://fishbase.org/)
[5] Integrated Taxonomic Information System (https://www.itis.gov/)
[6] International Standard Statistical Classification of Aquatic Animals and Plants (ISSCAAP) (https://www.fao.org/fishery/en/collection/asfis)

is reviewed and updated periodically to ensure high quality with valid classifications, revisions, and additions on newly described species.

4. Aquatic Species Knowledge Base Construction Process

Figure 1 shows how information was collected from the different data sources. More specifically, we used the CSV code lists for FAO ASFIS List of Species for Fishery Statistics Purposes[7], a RESTfull API for WoRMS[8] and FishBase[9] and the database dumb in MySQL from ITIS[10]. As illustrated in Figure 1, the resources collected from each source appear in different formats, so it is necessary to proceed with a syntax normalization phase before actually transforming them into instances of the top-level ontology MarineTLO. We started gradually integrating contents from the different data sources into the knowledge base, starting from FAO ASFIS, then WoRMS, then FishBase and the last one being ITIS. For each pair of data sources, we used different parts as the "glue" to connect species information. More specifically, we used the binomial of a species between FAO ASFIS and WoRMS, while for the rest of the pairs (i.e. WoRMS and FishBase, FishBase and ITIS) we used their codes and their scientific names. Finally, we should note that from ITIS, we collected only the marine species, based on the species that we already have added in the knowledge base until that moment. It is also worth noting, that we have collected all the possible external identifiers for species from WoRMS using the corresponding API method (i.e. AphiaExternalIDByAphiaID). These identifiers refer to IDs of the species in different data sources (such as GBIF[11], GISD[12], IUCN Red List[13], etc.) paving the way for including more data sources in future.

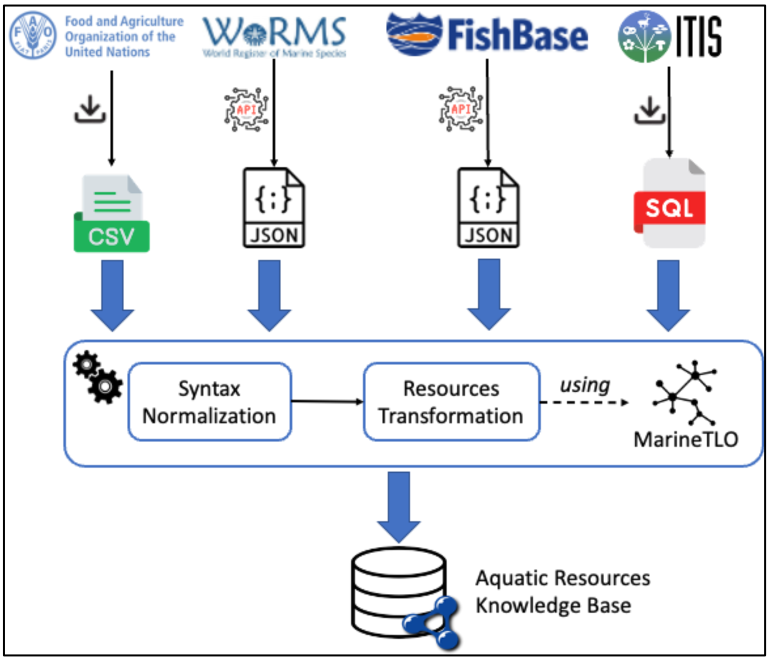

Figure 1. The process of collecting species resources from external data sources, preparing, adapting and ingesting them into the Knowledge Base.

During the construction of the knowledge base about species we have encountered several mismatches regarding the scientific names of species. To resolve those issues, we used as source of truth FishBase, that contains a list of synonyms for each species describing if a synonym is valid or not. For the few

[7] https://www.fao.org/fishery/en/collection/asfis
[8] https://www.marinespecies.org/rest/
[9] https://www.fishbase.ca/api/readme.txt
[10] https://www.itis.gov/downloads/index.html
[11] Global Biodiversirt Information Facility (https://www.gbif.org/)
[12] Global Invasive Species Database (https://www.iucngisd.org/gisd/)
[13] https://www.iucnredlist.org/

cases that the issue could not be resolved that way, the corresponding data source owners were informed about them. After communicating with them some of those issues were resolved. Figure 2 describes in a diagrammatic manner how the information is stored in the knowledge base by relying on the proper classes of MarineTLO, for example IDs are instances of the class MarineTLO:BC32_Identifier[14]. In this example, we report all the information that refer to the species with binomial "*Mullus barbatus*"; the left group of information report the different identifiers, the one in the middle report some of the names, and the right one report the taxonomic information of the species. Note that all the species-related information is presented in a uniform manner, although they have been collected from different data sources. Figure 3 is derived from a web application that allows searching the knowledge base and illustrates the available identifiers for marine species belonging to the genus "*Mullus*"

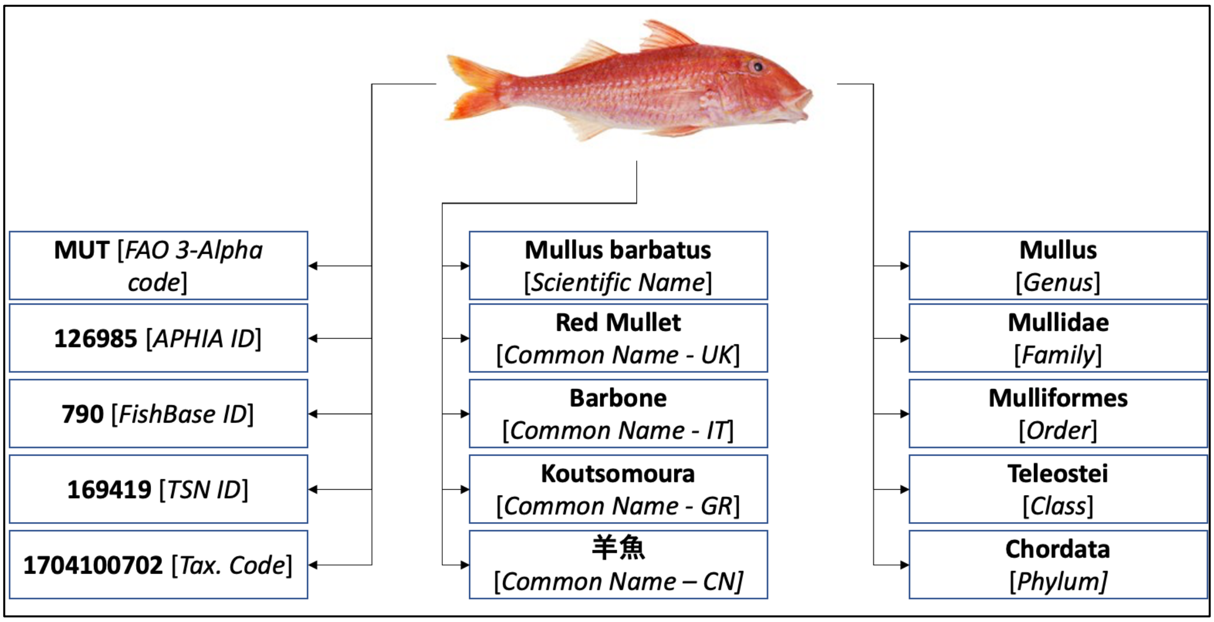

Figure 2. Detailed information (i.e. identifiers, names, taxonomic information) about aquatic species in the knowledge base

ASFIS ID	APHIA ID	FishBase ID	TSN ID	Taxonomic Code	Scientific Name
					Mullus
MUX				17041007XX	Mullus spp
	159418	FB-1093	169417		Mullus auratus
MWU	273661	FB-14063	620934	1704100703	Mullus argentinae
MUR	126986	FB-1327	169418	1704100701	Mullus surmuletus
	1525520	FB-25966			Mullus ponticus
MUT	126985	FB-790	169419	1704100702	Mullus barbatus

Figure 3. Identifiers of marine species belonging to the species "*Mullus*"

5. Aquatic Resources Knowledge Base

Of course, fisheries management is not just about aquatic species. It includes more essential information, like fishing assessment or management water areas and fishing gears. So, we applied the same methodology for constructing knowledge bases about those entities as well. As regards areas we mainly used FAO major fishing areas for statistical purposes[15], as well as Large Marine Ecosystems (LME) [9], Marine Regions[16], GFCM geographical subareas (GSA)[17], several national jurisdiction

[14] http://www.ics.forth.gr/isl/ontology/MarineTLO/BC32_Identifier
[15] https://www.fao.org/fishery/en/area/search
[16] https://www.marineregions.org/
[17] https://www.fao.org/gfcm/

areas and ISO3166 codes[10]. For fishing gears, we relied on the different versions of the FAO International Standard Statistical Classification of Fishing Gear (ISSCFG) standard [11].
Overall, we have implemented a process that constructs a knowledge base of aquatic resources from publicly available, well known data sources, focusing on aquatic species, water areas and fishing gears. The knowledge base can be browsed either through a SPARQL endpoint[18] or through a dedicated web application[19] that supports searching as well (an indicative screenshot is shown in Figure 3). Overall, it contains information about 40,564 marine species, 3,316 water areas, 88 fishing gear resources.
The Aquatic Resources Knowledge Base is actively used for monitoring the status of fisheries. In particular it is used by the Global Record of Stocks and Fisheries [12], for harmonizing and updating, if necessary, the information of the stocks and fisheries it collects.

6. Conclusion

In this paper, we describe the process for the construction of a comprehensive knowledge base for aquatic resources identifiers, to support the complex process of fisheries management. Although we focused on semantically integrating marine species, water areas and fishing gears, our methodology and tools can be applied for other entities or resources are well (e.g. identification of countries, fishery management authorities, etc.). The knowledge base can also be used as a reference towards harmonizing fishery management resources and we also report its current usage on this. Potential extensions include the addition of more data sources and entities as well as the automation of the construction process, so that the knowledge base remains in-sync with the contents of the data sources that are used.

7. Acknowledgements

This work has received funding from the Horizon Europe projects Blue-Cloud 2026 (Grant Agreement No: 101094227) and VeriFish (Grant Agreement No: 101156426).

8. References

[1] Ritchie, H. and Roser, M., 2023. Fish and overfishing. Our World in Data.
[2] Pikitch, E.K., Santora, C., Babcock, E.A., Bakun, A., Bonfil, R., Conover, D.O., Dayton, P., Doukakis, P., Fluharty, D., Heneman, B. and Houde, E.D., 2004. Ecosystem-based fishery management. Science, 305(5682), pp.346-347.
[3] Antoniou, G. and Van Harmelen, F., 2004. A semantic web primer. MIT press.
[4] De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A. and Rosati, R., 2018. Using ontologies for semantic data integration. A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, pp.187-202.
[5] Tzitzikas, Y., Allocca, C., Bekiari, C., Marketakis, Y., Fafalios, P., Doerr, M., Minadakis, N., Patkos, T. and Candela, L., 2013, November. Integrating heterogeneous and distributed information about marine species through a top level ontology. In Research conference on metadata and semantic research (pp. 289-301). Cham: Springer International Publishing.
[6] DiGiuseppe, N., Pouchard, L.C. and Noy, N.F., 2014. SWEET ontology coverage for earth system sciences. Earth Science Informatics, 7, pp.249-264.
[7] Schulz, S., Stenzhorn, H. and Boeker, M., 2008. The ontology of biological taxa. Bioinformatics, 24(13), pp.i313-i321.
[8] Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E. and Haendel, M.A., 2012. Uberon, an integrative multi-species anatomy ontology. Genome biology, 13(1), pp.1-20.
[9] Sherman, K. and Duda, A.M., 1999. Large marine ecosystems: an emerging paradigm for fishery sustainability. Fisheries, 24(12), pp.15-26.
[10] ISO 3166 Country Codes, URL: https://www.iso.org/iso-3166-country-codes.html

[18] https://isl.ics.forth.gr/grsf/sparql
[19] https://isl.ics.forth.gr/grsf/grsf-ir

[11] The International Standard Statistical Classification of Fishing Gear (ISSCFG). URL: https://www.fao.org/cwp-on-fishery-statistics/handbook/capture-fisheries-statistics/fishing-gear-classification/en/

[12] Marketakis, Y., Tzitzikas, Y., Gentile, A., Van Niekerk, B., and Taconet, M., 2020. On the Evolution of Semantic Warehouses: The Case of Global Record of Stocks and Fisheries. 14th International Conference on Metadata and Semantics Research, Special Track on Metadata & Semantics for Agriculture, Food & Environment (AgroSEM'20) Madrid, 2020.