BioOnto: Towards an Integration of Biological and Biogeographic Data

Marcos ZARATE [a,b,1], Agustina BUCCELLA [c] and Pablo FILLOTTRANI [d,e]

[a] *Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas*

[b] *LINVI, Faculty of Engineering, Universidad Nacional de la Patagonia San Juan Bosco*

[c] *GIISCO Research Group, Computer Science Department, Universidad Nacional del Comahue*

[d] *Computer Science and Engineering Department, Universidad Nacional del Sur*

[e] *Comisión de Investigaciones Científicas de la provincia de Buenos Aires*

Abstract In this work we present the preliminary design of *BioOnto*, an ontology-based system designed to integrate two important databases for Argentine science, the National System of Biological Data (SNDB) and the Ocean Biogeographic Information System (OBIS). *BioOnto* uses Web Ontology Language (OWL) to make assertions about classes in the underlying model and to define object properties that are used to link instances therein. To accomplish the integration, we follow the *Single Ontology Approach*. We illustrate the usefulness of *BioOnto* presenting fragments of the ontology that supports inference about (1) the relationship between preys and predators (2) which species coexist in a marine region (in particular the Argentine region is of our interest).

Keywords. Database Integration, Ontology, Biogeography, Biodiversity, Darwin Core

CEUR-WS.org/Vol-2050/ODLS_paper_6.pdf

1. Introduction

Biogeography [1] is a scientific discipline that studies the distribution of living beings on earth, as well as the processes that have originated it, that modify it and that can make it disappear. It is an interdisciplinary science, which is both a branch of geography and biology, receiving its foundations of specialties such as botany, zoology, ecology or evolutionary biology and other sciences such as geology. Currently there is a steadily growing wealth of biogeographic and biological data from a wide range of disciplines that being available from on-line information systems [2,3,4] around the world. In particular, as the number of species and geographic regions under study and also the participating institutions and research groups continue growing, addressing elaborate research issues about the complex interrelationships among these data becomes increasingly difficult. In the case that the information sources cannot be properly coordinated, data access and information discovery can sometimes become a daunting task. This drives the need

[1] Corresponding Author: (CESIMAR-CENPAT-CONICET), Boulevard Brown 2915 (U9120ACD), Puerto Madryn, Chubut, Argentina; E-mail: zarate@cenpat-conicet.gob.ar

to further facilitate the information integration across data sources, taking into account different querying interests in a diversity of fields and disciplines.

SNDB[2] supported by the Ministry of Science, Technology and Productive Innovation of Argentina and OBIS[3] supported by the Intergovernmental Oceanographic Commission of UNESCO, are currently two of main reference databases for ocean biologists, ecologists, and other researchers in Argentina. Both use Darwin Core Standard (DwC) [5] which is a general purpose vocabulary designed to facilitate the transfer and integration of biodiversity data. Though the DwC is defined in an RDF [6] document[4], integration of biodiversity data in the *Semantic Web* (SW) [7] is in its early stages. One of the major challenges for DwC in the SW context is the lack of a well-defined ontology. Without rigorous relationships between concepts and the properties that define them, connections between biodiversity data and related semantically rich information, such as literature and genomes, are difficult to traverse [8,9,10,11].

To facilitate the integration of these two databases, we present the design of *BioOnto*, an ontology that enables the integration following the *Single Ontology approaches* which uses one global ontology providing a shared vocabulary for the specification of the semantics and all information sources are related to one global ontology. The choice of this approach is due to the fact that it is ratified for many years as demonstrated in [12] its implementation is successful in cases where the nature of the data is similar. In addition, *BioOnto* is designed to link data together with other systems compliant with the RDF which follows the principles established by *Linked Open Data* initiative (LOD)[5].

The paper is organized as follows: Section 2 shows the state of art and the problems of integration that exist today. Section 3 describes the architecture, design considerations, and features of *BioOnto*. Section 4 presents a use case that demonstrates how the information from both data sets are integrated to better understand the behavior of a particular specie. Finally Section 5 presents the conclusions and the future work.

2. Related Work

In this section we will review the work related to the integration of biodiversity and biogeography domain. The first work consulted is [13] developed by OBIS-ENV-DATA, which proposes an extension of DwC to expand OBIS with environmental data to effectively manage combined datasets, although this generates advantages in the incorporation of environmental data, still persists the integration problem which has the DwC standard itself. In *MarineTLO* [14] the authors defined a core ontology for publishing marine data concerning to the European iMarine project[6], which is suitable for setting up warehouses that can serve complex queries. The main objective is to assemble information from different datasets to give more details of a particular marine species. However the DwC standard is not used to build the underlying ontology, limiting the possibility of integrating biodiversity data that respects the DwC standard. *BiSciCol* [15] describes an architecture to convert biodiversity data (in standard tabular formats) such as Darwin

[2] http://datos.sndb.mincyt.gob.ar/

[3] http://beta.iobis.org/

[4] http://rs.tdwg.org/dwc/rdf/dwcterms.rdf

[5] http://linkeddata.org/

[6] http://www.i-marine.eu/Pages/Home.aspx

Core-Archives [16] into RDF representations. In this case an ontology is described using the terms of the DwC but there are no examples of inferences or queries that may be of help to researchers interested in integrating different databases; on the other hand, currently the project page is not accessible. We plan to use DwC standard to capture complex aspect of biodiversity domain. In [17] the authors describe RDF-based data structures that are to be employed in the creation of a centralized repository of metadata from the heterogeneous data sources in the RITMARE research network[7]. A different approach is taken in [18] where the authors propose the creation of micro-linked open data clouds formed by oceanographic LOD-compliant datasets. A number of ontology design patterns called *GeoLink Modular Oceanography Ontology* are designed in [19] for the Oceanography domain. The resulting patterns are sufficiently modular, and thus arguably easier to extend than foundational top-level ontologies. Currently, the GeoLink project is in the middle of populating the patterns with actual data and a very preliminary evaluation demonstrated that the patterns together can serve as an integrating layer of heterogeneous oceanographic data repositories.

Current research is trying to address the existing gap to integrate biological, oceanographic and biogeographic data. However, critical look at the available literature indicates that there are limitations related to: (1) the absence of robust, standardized, and widely-accepted vocabularies and ontologies for linkable biodiversity and biogeography data. (2) the disagreements presently governing the use of identifiers in biodiversity and biogeography data is a major impediment to integrate these data.

3. Proposed Architecture and Ontology Details

The proposed architecture, showed in Figure 1, is divided into three layers: (1) **Input Data** representing the databases OBIS and SNDB in which the data can be downloaded in CSV format, then this information is converted to RDF using OpenRefine tool [20] where data inside these files are cleaned and converted to standardized data types (dates, numerical values, etc.). Data are converted to RDF triples using RDF Refine[8], which is an extension of OpenRefine. The columns of the CSV file are mapped as instances of DwC classes. Every resource must have a URI that can be used to link that resource to other resources both within this dataset and others anywhere on the web. The base URI that is common to the main classes is: `bio-onto:http://www.cenpat-conicet.gob.ar/ontology/`. In the CSV file we have a column with unique identifier, “ID”, to use as unique values in identity URIs. We use GREL (General Refine Expression Language[9]) to generate a new URI, the expression that we defined is `"occurrence/"+value.urlify()` this concatenates to the base URI the string `"ocurrence/"` along with the value of the ID, then a URI generated after applying the expression would look like this: `bio-onto:occurrence/valueID`. The resulting RDF to describe an occurrence is:

```
        SUBJECT                         PREDICATE              OBJECT
bio-onto:occurrence/f6bbf85d-85ea-4605  rdf:type           dwc:Occurrence
```

[7] http://www.ritmare.it/en/

[8] http://refine.deri.ie/

[9] https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions

`dwc:` is an abbreviation for the real namespace `http://rs.tdwg.org/dwc/terms/`, Table 1 describes the mapping performed together with the columns of CSV file used to generate the main URIs. The complete mapping can be consulted at[10]. (2) **Global Ontology and Storage** in which the RDF triples are stored in GraphDB[11]. This is a highly efficient and robust graph database with OWL inferences and SPARQL [21] support. Then the users can access the data resources easily using the different visualizations provided by GraphDB[12]. (3) **Information Retrieval** which provides a transparent interface to the non-expert user, so that they can perform queries without needing to know details of the SPARQL query language.

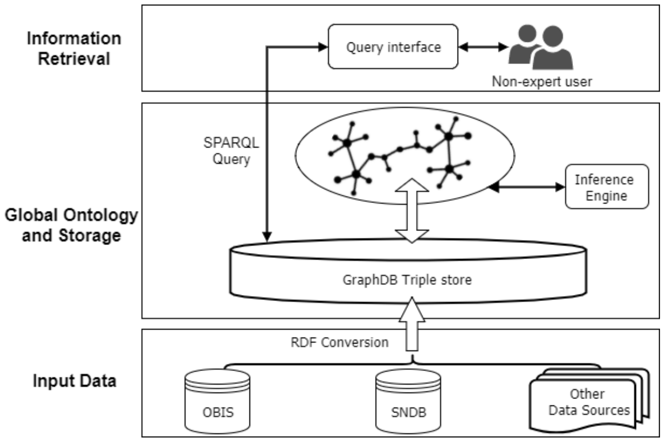

Figure 1. Proposed Architecture: the data of both datasets are converted to RDF, then the mapping corresponding to the classes defined in the ontology is performed. Finally the data can be queried through SPARQL queries.

3.1. A Multidisciplinary Approach

The development of the ontology is initiated within CENPAT[13]. The different research lines at CENPAT include marine biology, aquatic resource management, oceanography, paleontology, and biological diversity, among others. A working group at CENPAT is focused on working with issues related to the management and conservation of marine resources, this research group called CESIMAR acronym for (Center for the Study of Marine Systems) have a long tradition on studying the biology and behavior of marine mammals. Because most of the research done generates a large volume of data, but depending on its nature, they are published in SNDB (if they are biological data) or OBIS (if they are biogeographic data). Up to date there has not been developed a tool which allows integrating both databases. For example it is of great interest to know *which species coexist in a certain marine region*, or to establish *trophic relations between marine mammals*.

[10] `https://github.com/cenpat-gilia/BioOnto/blob/master/scripts/mapping.json`

[11] `http://www.ontotext.com/products/ontotext-graphdb/`

[12] `http://web.cenpat-conicet.gob.ar:7200/sparql`

[13] Patagonian National Research Centre (CENPAT), `http://www.cenpat-conicet.gob.ar/`

Table 1. The first part of the table shows the main classes corresponding to the categories of the DwC standard. You can also see the columns of the CSV file that were used to generate the corresponding URIs. The second part of the table shows the properties used and an example of the literals that are obtained from the columns of the file. For simplicity only the main ones are shown. `geo-pos:` is an abbreviation for the real namespace `https://www.w3.org/2003/01/geo/` and `foaf:http://xmlns.com/foaf/spec/` .

Class	Columns used to create URI	URI example
dwc:Taxon	genus + specificEpithet	<bio-onto:taxon/Mirounga_leonina>
dwc:Occurrence	id	<bio-onto:occurrence/f6bbf85d-85ea-4605>
dwc:Event	id	<bio-onto:event/f6bbf85d-85ea-4605>
dwc:Dataset	dataset	<bio-onto:dataset/dwca-mamcenpat-v1.1>
dc:Location	id	<bio-onto:location/f6bbf85d-85ea-4605>
foaf:Agent	institutionCode	<bio-onto:agent/cenpat-conicet>

Property	Columns used	Example
dwc:class	class	”Mammalia”∧∧xsd:string
dwc:family	family	”Phocidae”∧∧xsd:string
dwc:genus	genus	”Mirounga”∧∧xsd:string
dwc:kingdom	kingdom	”Animalia”∧∧xsd:string
dwc:order	order	”Carnivora”∧∧xsd:string
dwc:phylum	phylum	”Chordata”∧∧xsd:string
dwc:scientificName	scientificName	”Mirounga leonina”∧∧xsd:string
dwc:basisOfRecord	basisOfRecord	”PreservedSpecimen”∧∧xsd:string
dwc:occurrenceRemarks	occurrenceRemarks	”craneo completo”∧∧xsd:string
dwc:individualCount	individualCount	1∧∧xsd:int
dwc:CatalogNumber	CatalogNumber	”100751-1”∧∧xsd:string
geo-pos:lat	decimalLatitude	-42.53∧∧xsd:decimal
geo-pos:long	decimalLongitude	-63.6∧∧xsd:decimal
geo-ont:countryCode	country	”Argentina”∧∧xsd:string
dwc:verbatimEventDate	verbatimEventDate	”2004-10-22”∧∧xsd:date
foaf:name	recordedBy	”CENPAT-CONICET”@en .

After several interviews with the domain-experts, the key targets of the ontology model are specified to: (1) Establish marine regions, determined by maximum and minimum latitudes. (2) Establish which species of marine mammals are prey or predator and (3) Determine which species coexist in the economic areas pertaining to Argentina, in particular to see whether the commercial fishing does alter the ecological balance of the sea species.

3.2. Ontology Development Life-cycle

Regarding to the methodological strategy of our approach, we keep in line with the tradition that considers ontology as an engineering artifact that is useful to model some aspects of the world. That is why we adopted the methodology defined in [22]. The main stages are: (1) **Analysis:** We planned to use the ontology first with the DwC classes, we decided to first try and represent these objects as concepts and then see if some concepts were lacking or inadequate and eventually adjust the structure. (2) **Building the ontology:** The building process is iterative. Basically it can be broken down: finding conditions to constrain the concepts, introducing the properties and/or concepts needed to build the conditions and building the subsumption hierarchies of concepts and properties. (3) **Evaluation:** For the evaluation we take into account two important aspects highlighted in [23], *consistency* of the ontology since an inconsistent ontology would yield questionable results, this task is performed using *Ontology Debugger*[14] a plugin for Protégé and Pellet reasoner[15], we test the consistency after each set of changes we make, even if the changes are supposed to be simple. Another factor to evaluate is the *complexity*, one way to evaluate this is to ask the reasoner to classify the ontology. If this test takes too much time, it is likely that the ontology will not be usable in real conditions. If such is the case, corrections are to be made. Since usually the ontology size cannot be reduced, the general idea is to write simpler restrictions on properties. This means using a less complicated logic if possible. For instance, using existential restrictions instead of qualified cardinality restrictions helps keeping the complexity lower for the reasoner. (4) **Maintenance:** After complexity test is performed with adequate performance, we check the ontology's performance in real use. This is done by testing the applications exploiting the ontology and evaluating the performance, both in terms of execution speed and results quality. The analysis of the results help us fine tune the ontology to our exact needs.

3.3. The Ontology Structure

At this point we might think that the integration of both databases is solved since both are converted to RDF, their integration should be simple given that a URI can indicate what kind of entity it identifies. But this is not enough, it is necessary to impose a layer of computationally tractable meaning, where the relationships that hold between them can be accurately interpreted and used in an automated way. For example, establishing relationships between instances of classes through an inverse property, defining cardinality restrictions or establish subsumption relationships. Description Logics (DL) [24] is an adequate means of representing ontologies. Furthermore, OWL is based on DL, so we decided to describe our ontology using DL and to implement it in OWL-DL [25]. Both of these flavors are well-supported by existing reasoners and is particularly suitable for the type of reasoning we intend to perform at this stage of the work. *BioOnto* is based primarily on the DwC classes and terms, but also classes and properties defined by domain experts.

As we mentioned earlier DwC [5] is a body of standards for biodiversity informatics. It provides stable terms and vocabularies for sharing biodiversity data. DwC is maintained by TDWG[16] (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases). These terms are organized into nine categories (often referred to as *classes*), six of which cover broad aspects of the biodiversity domain. *Occurrence* refers to existence of an organism at a particular place at a particular time, *Location* is the place where the organism were observed (normally a geographical region or place), the *Event* class is the relationship between *Occurrence* and *Location* this registers sampling protocols and methods, dates, time and field notes, *Taxon* refers to scientific names, vernacular names, etc. of the organism observed. The remaining categories cover relationships to other resources, measurements, and generic information

[14] `https://git-ainf.aau.at/interactive-KB-debugging/debugger/wikis/onto-debugger`

[15] `https://github.com/stardog-union/pellet`

[16] `http://www.tdwg.org/`

about records which in our case are not used. Specifically for the record level, DwC recommends the use of a number of terms from Dublin Core[17].

3.3.1. Classes and Properties

The main classes of the ontology are taken from the vocabulary specified in DwC, as well as some of classes and properties therein. The *BioOnto* structure and properties assigned to its classes provide a series of useful reasoning tasks that can be formed by SPARQL queries (see Section 4 for examples). The general model of *BioOnto* can be seen in Figure 2. The class hierarchy of the OWL ontology consists of seven key classes: **Occurrence** represents instances where the presence of an organism at a *Location* is observed. It functions primarily as a node that connects *Taxon* to *Events*. **Location** is a spatial region or named place. For DwC, a set of terms describing a place, whether named or not. **Event** functions primarily as a node to connect one or more *Occurrences* to an *Event*, and one or more *Events* at a *Location*. **Taxon** represents a unit of biodiversity e.g. species, genus, family, etc. **Agent** *foaf:Agent* class is imported directly into *BioOnto*. An agent (e.g. person, group, software or physical artifact). **Region** is an area of the ocean delimited by a latitude and longitude. **Dataset** is an identification of the data set belonging to OBIS or SNDB.

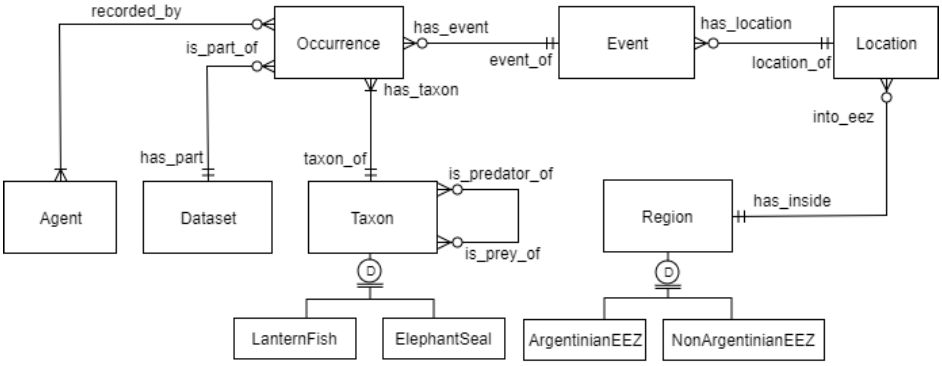

Figure 2. Entity-relationship diagram of main classes and relationships defined in BioOnto.

BioOnto defines a number of properties used to link classes (see Table 2), and also reuses some properties of some commonly used terms like Basic Geo Vocabulary (WGS84 lat/long[18]) and FOAF[19]. In this way *BioOnto* accrues most of the scientific and technical vocabulary required to achieve a semantic understanding of the most commonly used terms in biogeography and biodiversity, the ontology in OWL can be seen in[20].

[17] http://dublincore.org/documents/dcmi-terms/

[18] https://www.w3.org/2003/01/geo/

[19] http://xmlns.com/foaf/spec/

[20] https://github.com/cenpat-gilia/BioOnto/blob/master/ontology/BioOnto.owl

Table 2. Object properties that link the main classes in the BioOnto model. Some properties like is_predator_of has an inverse property linked by an owl:inverseOf relationship. For each property, the intended subject class is declared the rdfs:domain of the term and the intended object class is declared the rdfs:range of the term.

Object Property	Domain	Range	Inverse
has_event	Occurrence	Event	event_of
has_location	Event	Location	location_of
has_taxon	Occurrence	Taxon	taxon_of
into_eez	Location	Region	has_inside
part_of	Occurrence	Dataset	has_part
is_predator_of	Taxon	Taxon	is_prey_of
recorded_by	Occurrence	Agent	-

4. Case study: Behavior of Marine Mammals

This section develops use cases that are relevant to researchers dedicated to the conservation of the marine species and the study of animal behavior. One of the most common problems faced by the scientists is to determine which species coexist in the *Argentinean Exclusive Economic Zone* (AEEZ[21]). One of the species that is of particular interest to scientists in the Argentine sea littoral is *Mirounga Leonina* or (Southern Elephant Seal), because of the relevant information that provides with respect to the sea. For this reason, *Mirounga Leonina* are used as oceanographic sampling platforms [26], they are ideal carriers of electronic devices that register parameters like salinity, position, depth and temperature, providing huge amounts of information associated with the key habitats.

The data collected by different individuals during their migration is captured by these devices, and are available in OBIS. In some cases, new biodiversity related knowledge may be discovered by the scientists using SNDB, (*e.g. determine which fish species live in the same region than the elephant seals*). This information may not exist in OBIS records, and therefore it is highly desirable to be able to integrate this information from SNDB in a transparent way. The following subsections show examples where the axioms that were defined in the ontology to enable them to answer these questions.

4.1. Inferring Predator/Prey Relationship

Scientists recently discovered that a large percentage of Southern Elephant Seal feeding is the *Lanternfish* [27]. Then it is possible to define axioms in OWL to establish the relationship predator/prey, a knowledge that was only implicit in the ontology. Using rolification and the chain axiom property it can be expressed the above inference in OWL with the following three axioms expressed in Manchester OWL [28] syntax:

```
(1) R_Elephant some Self
(2) R_Lanternfish some Self
(3) R_Elephant o owl:topObjectProperty o R_Lanternfish SubPropertyOf
    is_predator_of
```

[21] http://www.marineregions.org/gazetteer.php?p=details&id=8466

Where R_Elephant and R_LanternFish are object properties, and it is required to add atoms of the form U(t,u), where U is the universal property, (*i.e.*, owl:topObjectProperty). In this way after executing the reasoner, all instances of the class ElephantSeal will be related through the property is_predator_of with all instances of the class LanternFish. This approach was taken from [29] where instead of using SWRL to define rules, we use OWL axioms.

4.2. Inferring Species into Argentinean Exclusive Economic Zone

To determine the species that inhabit the AEEZ, we can use (since the data are georeferenced by latitude and longitude) axioms OWL to infer that a certain specie is present in the AEEZ, for this we use the object property into_eez and the following axioms expressed in Manchester OWL syntax.

```
(4) Location and (Lat some xsd:double[>="-58.4046"^^xsd:double,
    <="-32.4483" ^^xsd:double]) and (long some xsd:double
    [>="-69.6095"^^xsd:double,<= "-52.631" ^^xsd:double])
    SubClassOf into_eez value AEEZ
```

Any instance of the Location class that is within the maximum and minimum values of latitude and longitude, will be related by the property into_eez with an instance of ArgentineanEEZ class called AEEZ which represents the exclusive economic zone of Argentina.

4.3. Querying and Reasoning Facilitated by BioOnto

This section shows an example of simple SPARQL queries that allow to explore the data of both datasets, in particular we will make use of the information associated to the species as well as to their location. The following query allows to retrieve the location of all mammals containing the two datasets. This particular query is important because it allows in the future to work with interfaces that allow the visualization of georeferenced data. To run this query through the SPARQL interface of GraphDB, see[22].

```
PREFIX geo-pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX bio-onto:<http://www.cenpat-conicet.gob.ar/ontology/>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>

SELECT ?s_name ?lat ?long
WHERE {
        ?s a dwc:Occurrence.
        ?s bio-onto:has_taxon ?taxon.
        ?taxon dwc:scientificName ?s_name.
        ?taxon dwc:class ?class.
        ?s bio-onto:has_event ?event.
        ?event bio-onto:has_location ?loc.
        ?loc geo-pos:lat  ?lat .
        ?loc geo-pos:long ?long .
```

[22] http://web.cenpat-conicet.gob.ar:7200/sparql?savedQueryName=bio-Loc-of-mammals

```
        FILTER regex(STR(?class), "Mammalia")
}
```

In this section we show examples that respond to the questions raised by researchers studying animal behavior. In the case of predator/prey relations is interesting because it allows to determine by reasoning, the species that are part of the food chain obtaining the species of OBIS and SNDB. In addition to this reasoning, we can determine when a species is considered a *top predator*, if it is not prey to any other species. Regarding to the marine regions, we offer the possibility that the user can define a specific marine region to perform some type of analysis. Reasoning allows to identify the species that were observed there and determine, for example if the existence of a specie in that region is due to the fact that it is feeding from another one.

5. Conclusions and Future Work

We have presented the initial steps performed in the ontology design called *BioOnto* developed with researchers from CENPAT. The proposed model enables interoperability and a common knowledge representation among databases *SNDB* and *OBIS*, allowing the retrieval of information that cannot be gathered by any of the individual information sources alone. Specifically, we presented the preliminary results of the ontology that (1) allows establishing predator/prey relationships between marine species, (2) define marine geographic regions and determine what species live there.

We are currently working on the integration of other databases (*e.g. FishBase*[23]) which are required in research tasks within the CENPAT to answer questions about the balance of species and their relationship with commercial fishing.

We know that this databases could be integrated into the **input data** layer since the proposed architecture is flexible to integrate data in structured or semi-structured format. Furthermore in future works we will consider to link our ontology with relevant marine ontologies[24].

References

[1] Mark V Lomolino and James H Brown. *Biogeography*. Number QH84 L65 2006.

[2] PN Halpin, Andrew J Read, BD Best, KD Hyrenbach, E Fujioka, MS Coyne, Larry B Crowder, SA Freeman, and C Spoerri. Obis-seamap: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles. *Marine Ecology Progress Series*, 316:239–246, 2006.

[3] Vishwas Chavan and Lyubomir Penev. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, 12(15):S2, 2011.

[4] Vishwas S Chavan and Peter Ingwersen. Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC bioinformatics*, 10(14):S2, 2009.

[5] John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, 2012.

[23] http://www.fishbase.org/

[24] http://mmisw.org/

[6] Ora Lassila and Ralph R Swick. Resource description framework (rdf) model and syntax specification. 1999.

[7] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[8] Steven J Baskauf, John Wieczorek, John Deck, and Campbell O Webb. Lessons Learned from Adapting the Darwin Core Vocabulary Standard for Use in RDF.

[9] Benjamin M Good and Mark D Wilkinson. The life sciences semantic web is full of creeps! *Briefings in bioinformatics*, 7(3):275–286, 2006.

[10] O James Reichman, Matthew B Jones, and Mark P Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018):703–705, 2011.

[11] Anne E Thessen and David J Patterson. Data issues in the life sciences. 2011.

[12] Yigal Arens, Chun-Nan Hsu, and Craig A Knoblock. Query processing in the sims information mediator. *Advanced Planning Technology*, 32:78–93, 1996.

[13] Daphnis De Pooter, Ward Appeltans, Nicolas Bailly, Sky Bristol, Klaas Deneudt, Menashè Eliezer, Ei Fujioka, Alessandra Giorgetti, Philip Goldstein, Mirtha Lewis, et al. Toward a new data standard for combined marine biological and environmental datasets-expanding obis beyond species occurrences. *Biodiversity Data Journal*, (5), 2017.

[14] Yannis Tzitzikas, Carlo Allocca, Chryssoula Bekiari, Yannis Marketakis, Pavlos Fafalios, Martin Doerr, Nikos Minadakis, Theodore Patkos, and Leonardo Candela. Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology. In *Metadata and Semantics Research: 7th Research Conference, MTSR 2013, Thessaloniki, Greece, November 19-22, 2013. Proceedings*, pages 289–301. Springer International Publishing, 2013.

[15] Brian J Stucky, John Deck, Tom Conlin, Lukasz Ziemba, Nico Cellinese, and Robert Guralnick. The biscicol triplifier: bringing biodiversity data to the semantic web. *BMC bioinformatics*, 15(1):257, 2014.

[16] K Döring M Robertson T Remsen D, Braak. Darwin Core Archive How-To Guide. 2011.

[17] Cristiano Fugazza, Anna Basoni, Stefano Menegon, Alessandro Oggioni, Fabio Pavesi, Monica Pepe, Alessandro Sarretta, and Paola Carrara. RITMARE: Semantics- Aware harmonisation of data in Italian marine research. In *Procedia Computer Science*, 2014.

[18] Adam Leadbetter, Robert Arko, Cynthia Chandler, Adam Shepherd, and Roy Lowry. Linked Data An Oceanographic Perspective. *The Journal of ocean Technology*, 8(3), 2013.

[19] Adila Krisnadhi, Yingjie Hu, Krzysztof Janowicz, Pascal Hitzler, Robert Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Timothy Finin, Peng Ji, Matthew Jones, Nazifa Karima, Kerstin Lehnert, Audrey Mickle, Thomas Narock, Margaret OBrien, Lisa Raymond, Adam Shepherd, Mark Schildhauer, and Peter Wiebe. The GeoLink modular oceanography ontology. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[20] Ruben Verborgh and Max De Wilde. *Using OpenRefine*. Packt Publishing Ltd, 2013.

[21] Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF – W3C recommendation. Technical report, W3C, 2008.

[22] Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[23] Steffen Staab and Rudi Studer. *Handbook on ontologies*. Springer Science & Business Media, 2010.

[24] Franz Baader. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[25] Ian Horrocks, Peter F Patel-Schneider, and Frank Van Harmelen. From shiq and rdf to owl: The making of a web ontology language. *Web semantics: science, services and agents on the World Wide Web*, 1(1):7–26, 2003.

[26] Fabien Roquet, Carl Wunsch, Gael Forget, Patrick Heimbach, Christophe Guinet, Gilles Reverdin, Jean Benoit Charrassin, Frederic Bailleul, Daniel P. Costa, Luis A. Huckstadt, Kimberly T. Goetz, Kit M. Kovacs, Christian Lydersen, Martin Biuw, Ole A. Nøst, Horst Bornemann, Joachim Ploetz, Marthan N. Bester, Trevor McIntyre, Monica C. Muelbert, Mark A. Hindell, Clive R. McMahon, Guy Williams, Robert Harcourt, Iain C. Field, Leon Chafik, Keith W. Nicholls, Lars Boehme, and Mike A. Fedak. Estimates of the Southern Ocean general circulation improved by animal-borne instruments. *Geophysical Research Letters*, 2013.

[27] Jade Vacquié-Garcia, Christophe Guinet, Cécile Laurent, and Frédéric Bailleul. Delineation of the southern elephant seal s main foraging environments defined by temperature and light conditions. *Deep Sea*

Research Part II: Topical Studies in Oceanography, 113:145–153, 2015.

[28] Matthew Horridge, Nick Drummond, John Goodwin, Alan L Rector, Robert Stevens, and Hai Wang. The manchester owl syntax. In *OWLed*, volume 216, 2006.

[29] Md Kamruzzaman Sarker, David Carral, Adila Alfa Krisnadhi, and Pascal Hitzler. Modeling owl with rules: The rowl protege plugin. In *International Semantic Web Conference (Posters & Demos)*, 2016.