Democratizing Argentine Marine Science Data Through Linked Open Data

Marcos Zárate[1,2,*] [0000-0001-8851-8602] Carlos Buckle[2] [0000-0003-0722-0949], Mirtha Lewis[1,3] [0000-0001-6262-6226], Claudio Delrieux[4] [0000-0002-2727-8374], Dario Ceballos[1], and Gustavo Nuñez[2]

[1] Centre for the Study of Marine Systems, Patagonian National Research Centre (CESIMAR-CENPAT-CONICET), Puerto Madryn, Argentina.
zarate@cenpat-conicet.gob.ar

[2] Laboratorio de Investigación en Informática (LINVI) - Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Puerto Madryn, Argentina.
cbuckle@unpata.edu.ar

[3] Centro de Investigaciones y Transferencia Golfo San Jorge, (CIT-GSJ-CONICET), Comodoro Rivadavia, Argentina.
mirtha@cenpat-conicet.gob.ar

[4] Computer Science and Engineering Department, Universidad Nacional del Sur (DIEC-UNS), Bahia Blanca, Argentina
cad@uns.edu.ar

Abstract In this paper we expose experiences carried out during the last five years in the domain of Argentine marine sciences. Specifically data generated by Pampa Azul Argentine initiative to improve the publication of data using the advantages provided by Linked Open Data (LOD), Knowledge Graph (KG) and FAIR principles. The focus is on: a) to provide a conceptual analysis of traditional data publication in marine science, b) to describe projects based on LOD that involve information from Argentina, we mainly focus on the OceanGraph KG project, c) generate recommendations for data management for its best use in marine science.

Keywords: Linked Open Data · FAIR · Pampa Azul · Knowledge Graph · Marine Science.

1 Introduction and Motivation

In July 2020, Pampa Azul initiative was relaunched [1] aimed at promoting scientific knowledge, technological development and productive innovation in the South Atlantic Ocean, in order to develop a culture of the sea in Argentine society, promote the sustainable use of marine natural assets and strengthen the growth of the associated national industry.

* Corresponding author.

However, data management of data generated during first launch of Pampa Azul (2014) made it clear that data management and modeling of online data needed to be planned, safeguarded and shared, so products generated can be used by the scientific community for an adequate understanding of the functioning of our marine spaces. Furthermore, integration of these data with global federated databases was not taken into account, which makes their comprehensive scientific use very difficult.

Prior to Pampa Azul, the Ministry of Science, Technology and Productive Innovation developed the National Biological Data System [2] (SNDB by its acronym in Spanish) on a platform called Integrated Publishing Toolkit (IPT) [3] developed by Global Biodiversity Information Facility (GBIF) for biological data based on the Darwin Core standard [4] which is widely adopted by the international community. For marine data, the National Sea Data System (SNDM by its acronym in Spanish) was created on the platform developed by the International Oceanographic Data and Information Exchange (IODE), in order to visualize the information of the national satellite-type oceanographic data producing centers of Argentina.

From political and scientific contexts, a demand for data management is identified within the context of Pampa Azul, but the platform used by the SNDM provides information from the different resources in a fragmented manner and there are not sufficiently detailed resources for data entry. In addition, it does not allow integration with other scientific data repositories, this generated a lack of interest in contributing data sets. In 2018 only three new data sets were recorded and in 2019 there were no new data. These difficulties in data management are because they involve conceptual frameworks from different disciplines, such as Oceanography (physical, chemical and biological), Geosciences, and Meteorology, among which there is a great diversity in the types and formats of data to be managed. Nowadays, this data portal is down (see: https://www.argentina.gob.ar/ciencia/sistemasnacionales/datos-del-mar), among other things because it does not allow interactive viewing of data or other types of information, making it an unattractive tool for researchers or the general public interested in oceanography. Figure 1 shows the timeline with the most important milestones in marine science data management at local level.

Several causes can be mentioned that led to the failure of the SNDM, but we consider that the most important is due to the fact that marine sciences generate large volumes of data, due to advances in remote acquisition technology and the permanent emergence of new oceanographic campaigns [5]. Thus, it is necessary to develop systems capable of managing their integration and communication, both for comprehensive and secondary use by the participating groups and institutions, as well as for external users who require information.

One of the promising approaches to address the problems associated with heterogeneous data management is to store the information in a structured way and to represent data sets as graphs [6] which has been used in research and business, generally in close association with Semantic Web technologies [7], Linked Open Data (LOD) [8], large-scale data analytics, and cloud computing.

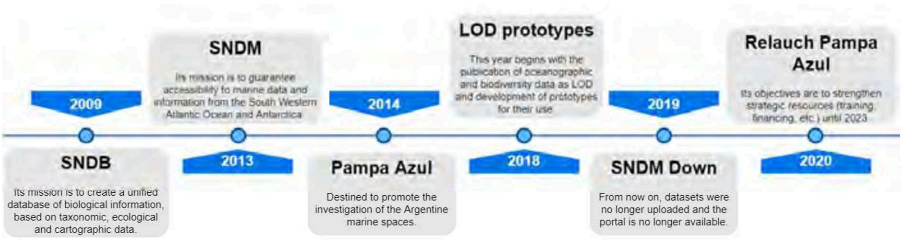

Fig. 1. Relevant milestones on marine data management in Argentina.

The main contribution of this paper is to focus on how LOD contributed the opening and democratization of marine science information, which is financed with public funds and show the lessons learned, so that future developments take into consideration from the beginning the opening of the data, complying with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) [9] for its best use.

The remainder of this paper is structured as follows: Section 2, presents details of the main developments using LOD in the field of Argentine marine sciences. Section 3, enumerates preliminary results obtained which can be reused in future developments. In Section 4, we discuss the principles of LOD that tie everything together. Finally, in Section 5, we present some lessons learned in these years to serve as experience in future research related to marine science.

2 LOD prototypes

Taking into account the limitations of the conventional systems described in Section 1, we have developed different proofs of concept in the oceanographic domain and marine biodiversity, which were reported in different scientific journals and conferences. We summarize these efforts below.

2.1 Linked Data in oceanography

Regarding the management and modeling of Argentine marine science data such as KG and LOD we can enumerate:

- 2018: The first development was based on the publication of metadata from oceanographic campaigns related to Pampa Azul [10].
- 2019: an oceanographic KG prototype called OceanGraph was defined in [11]. It is currently under development integrating new data sources. A simplified view of the proposed KG is shown in Figure 2.
- 2020: in [12] the potential uses of OceanGraph were demonstrated with a concrete example by specialists.
- 2022: a LOD dataset of observational data and hydrographic profiles of the South Atlantic Ocean was published as LOD in [13]. This data set was integrated into the structure of OceanGraph KG.

Based on the experience gained in these works, a series of recommendations related to the interoperability and integration of information from the Global Ocean Observing System (IOOS) were published in [14], this work being carried out in collaboration with the U.S. Integrated Ocean Observing System, the Norwegian Institute of Marine Research (IMR) and the National Centers for Environmental Information (NCEI).

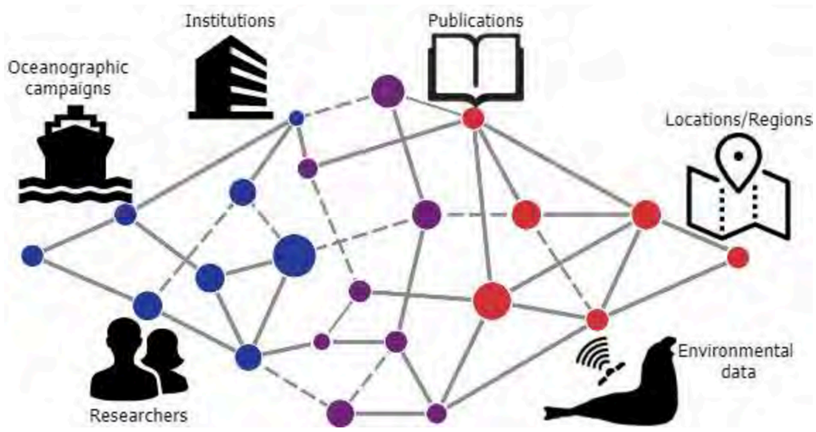

Fig. 2. Overview of *OceanGraph*: integrates information from oceanographic campaigns, scientific publications, researchers, institutions, marine locations and regions, and environmental variables from sensors placed on animals.

2.2 Linked Data in marine biodiversity

In addition to previous developments, we have published a marine biodiversity LOD dataset [15] and developed a linked data dashboard to visualize and complement information on certain species with other linked datasets [16]. In [15], we collaborate with Research Group on Languages and Artificial Intelligence (GILIA-UNCOMA) of Universidad Nacional del Comahue and Department of Computer Science and Engineering (DCIC) of Universidad Nacional del Sur, while [16], was a collaboration with the Argentine node of the Global Biodiversity Information Facility (GBIF)[5] and VertNet[6].

Finally, an ontology-based system called BiGe-Onto [17] was developed for the integrated management of marine Biodiversity and Biogeography information and a LOD dataset publicly available through DOI 10.5281/zenodo.3235548.

[5] `https://www.gbif.org/es/country/AR/summary`
[6] `http://vertnet.org/`

3 Preliminary results

As we detailed in the previous section, since 2018 we have developed different proposals using ontologies, LOD and publication of oceanographic data following the FAIR principles, in this section we discuss the preliminary results obtained and how these developments can be reused by other initiatives whose data pertain to marine sciences.

From a theoretical point of view, a survey and selection of standard ontologies was carried out for the extension of BiGe-Onto, which will later be the central model of OceanGraph. The survey was carried out from public repositories of ontologies and the subsequent analysis of the conceptual models underlying said ontologies. From a practical point of view, a web application is being implemented to present metadata results visually on data from oceanographic surveys, types of sampling carried out, people and institutions involved, and recorded environmental variables. This implementation requires the design of a software architecture and the subsequent selection of current web development and semantic web technologies. In this last group, we work on the selection of knowledge graph storage systems that also provide efficient search engines. A logical reasoner is being developed for data validation with a temporal dimension that allows the modeling of data in temporal logic from relational databases. Also, work is being done on the validation and scalability of this approach with case studies from different domains. As a summary, we can list the following results:

- a network of ontologies (semantic model), modeling marine science domain, with focus on sensor data and biodiversity.
- a LOD dataset.
- a proof-of-concept application to explore visually the KG.
- a set of running examples that potential consumers can use as training material. They consist of natural language Competency Questions (CQs) and their corresponding SPARQL queries [18].
- a SPARQL endpoint[7] to explore the resource, run tests, etc.

Regarding the semantic model we consider that the main contribution to highlight it is the main component of OceanGraph is a KG, intended as the union of the ontology network defined in Web Ontology Language (OWL) [19] and LOD data. Nevertheless, the KG is released as part of a package including accompanying material (documentation and online services) that support its consumption, understanding and reuse. OceanGraph bases its main structure on the relationships established between the selected datasets. The main classes that we define and reuse are: *campaigns, occurrences, papers, researchers, environmental variables and positions.* If a researcher consults OceanGraph, the expected results could recover one or more oceanographic campaigns in which she/he was involved from SNDM, datasets they collected from GBIF and Ocean

[7] `https://linkeddata.cenpat-conicet.gob.ar/snorql/`

Biogeographic Information System (OBIS)[8], and papers written by themself (from Springer Nature SciGraph)[9].

In the same way, the user could query data related to the occurrence of a species and the KG must retrieve in which campaigns it was observed, the information of the person who collected it, the exact place and date and associated variables that may be of importance (*e.g.*, weather or other environmental conditions during the collection).

We reuse only the elements from these ontologies that are necessary for modeling our data, adopting a *soft reuse* strategy [20] instead of importing the whole ontologies. OceanGraph ontology network consists of several ontologies modules connected by `owl:imports` axioms (See Figure 3). A list of prefixes and their corresponding URIs are listed in Table 1.

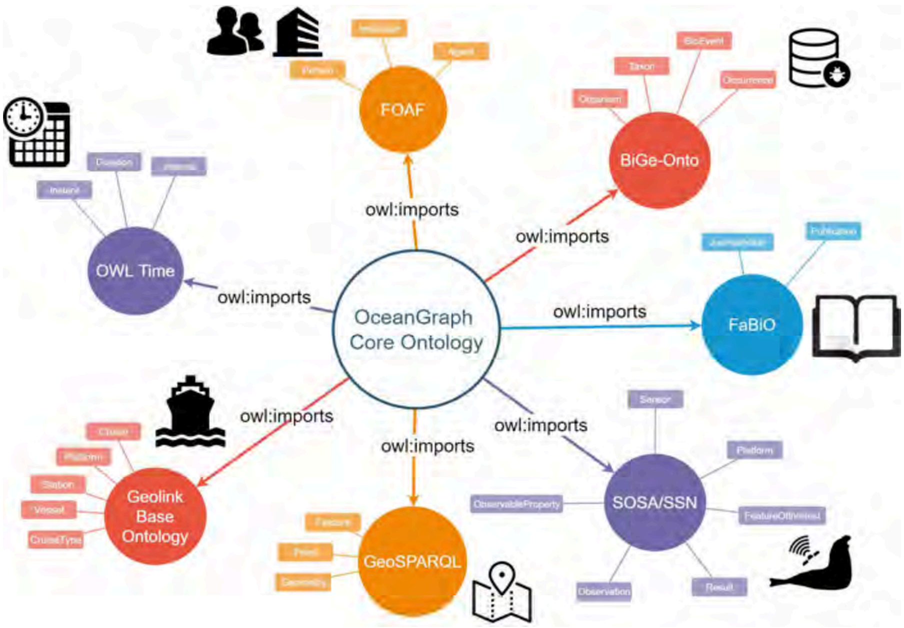

Fig. 3. OceanGraph ontology network intended to be adapted to different domains and reused by different marine science projects.

To synthesize the results obtained, we can highlight that the semantic model is generic enough to incorporate new data sources and be reused in other projects. Compliance with the FAIR principles allows the information from Argentine marine sciences to be visible and reusable by third-party applications that are interested in its exploitation.

[8] http://www.iobis.org/

[9] https://www.springernature.com/gp/researchers/scigraph

Table 1: Reused vocabularies and ontologies.

Ontology/Vocabulary name	Prefix
BiGe-Onto ontology	bigeonto
Semantic Sensor Network Ontology	ssn
Sensor, Observation, Sample, and Actuator Ontology	sosa
Darwin Core (literal values)	dwc
Darwin Core (IRI values)	dwciri
GeoSPARQL ontology	geosparql
W3C Time Ontology	time
FRBR-aligned Bibliographic Ontology	fabio
NERC vocabulary server (measured phenomena)	P01
NERC vocabulary server (biological entity sex)	S10
Quantities, Units, Dimensions and Types Ontology (v1.1) vocabulary	qudt
Quantities, Units, Dimensions and Types Ontology (version 1.1) schema	qudts
GoodRelations (v1.0)	gr
Simple Knowledge Organization System	skos

4 Discussion

In this section we discuss the aspects related to fulfilment of LOD principles of and aspects related to the semantic model used in OceanGraph.

LOD [21] is an idea from the Semantic Web [7] aimed at ensuring that data published on the Web is reusable, discoverable, and more importantly, that data published by different entities can work together. LOD principles are summarized in:

- Use URIs as names for things.
- Use HTTP URIs so people can look up these things.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
- Include links to other URIs so they can discover more things.

We have followed these guidelines when creating all datasets described in Section 2. Below we discuss each of these points separately.

Usage of URIs as resource identifiers Each instance is uniquely identifiable by an HTTP URI. For example, we define the result of a measurement of the average depth of the water column measured by an instrument as: `http://linkeddata.cenpat-conicet.gob.ar/data/result/id-233/avg_depth`. All instance identifiers follow this scheme.

Usage of HTTP URIs and dereferencing : According with linked data principles, we use dereferenceable HTTP URIs for our resources. For example for the average depth URI above, we generate a human-readable version of the dereferenced version using the URL: `http://linkeddata.cenpat-conicet.gob.ar/page/result/id-233/avg_depth` to dereference the URI.

Linking to other resources All resources in OceanGraph form a graph (there are no disconnected parts). In addition, resources are linked to external databases via properties like `owl:sameAs`, `skos:broader` and `skos:exactMatch`. These identifiers can be: ORCIDs, Wikidata entities, DBPedia resources, NERC vocabulary server ID´s, etc. We have created links between people and their ORCID records, publications and their OpenCitations records, as well as the environmental variables were related to the identifier in NERC. For example, average water temperature corresponds to the NERC identifier `SDN:P02::TEMP` through `skos:broader` property. See: `http://linkeddata.cenpat-conicet.gob.ar/resource/observableProperty/id-233/avg_temp` to understand this implementation.

Availability, sustainability, and licensing One of the most important design decisions when developing a KG is the platform that supports it. After several performance comparisons, we decided to use GraphDB[10] since it allows a quick integration of new sources of information, analyzes structured data in CSV, XLS, JSON, XML or other formats, it allows to generate data in RDF and store it in a local or remote SPARQL endpoint, and last but not least, it allows to clean the input data with a generic script language. GraphDB allows users to explore the hierarchy of RDF classes and its instances (Class hierarchy menu). In the same way, we can check the relationships between the KG classes and visually explore how many links were created between different class instances (Class relationship). To access the OceanGraph dataset, the user must authenticate themselves on `http://web.cenpat-conicet.gob.ar:7200/login`, using the following credentials (user: **oceangraph** password: **ocean.user**). *OceanGraph KG* is also available for download in DOI: 10.17632/9t5xkt9wwk.1 under CC BY 4.0 license.

5 Conclusions

Tim Berner-Lee suggested LOD principles [7] for judging data quality by its accessibility (open data access), by its format and structures, and by its interoperability with other data sources. The FAIR data principles have been introduced for similar reasons with a greater emphasis on achieving reuse. LOD gives a clear mandate to the opening of the data, while FAIR requires an established license for access and therefore includes the concept of reuse under consideration in the license agreement. In addition, FAIR makes a strong reference to contextual information required to improve data reuse. In accordance with LOD principles, such metadata would be considered interoperable data as well, however the requirement to augment the data with metadata indicates that FAIR is an extension of the LOD [22]. Our recommendation based on mistakes mentioned in Section 1 for data management in marine sciences is: it is not enough to develop useful applications for specific users, conception of these applications

[10] `http://graphdb.ontotext.com/`

must contemplate compliance with the FAIR principles so that they are truly useful. In particular the use of LOD from the beginning, this facilitates reuse by scientists and non-expert users, on the other hand it facilitates interoperability with other systems allowing more complex analyses.

As explained in Section 4, publication of the LOD version of OceanGraph allows compliance with a large part of the FAIR principles. There is a description of the data online, the data is available as RDF, and there are many links to structured vocabularies, and metadata about the collection is made available.

We envision OceanGraph as an integral part of the existing semantic network of marine science knowledge in Argentina, based on HTTP identifiers and controlled vocabularies. By enhancing and semantically linking OceanGraph knowledge to existing machine-readable data, we increase the quality of marine science data and increase the potential for reuse.

Acknowledgments: This research received funding from project *Linked Open Data Platform for Management and Visualization of Primary Data in Marine Science.* Project No. PI-1562. Financed by Secretariat of Science and Technology of the National University of Patagonia San Juan Bosco (UNPSJB).

References

1. Se relanzo la iniciativa pampa azul. https://www.argentina.gob.ar/noticias/se-relanzo-la-iniciativa-pampa-azul, 2020. [Online; accessed 23-Feb-2022].
2. Portal de datos de biodiversidad argentina. https://datos.sndb.mincyt.gob.ar/, 2014. [Online; accessed 23-Mar-2022].
3. Tim Robertson, Markus Döring, Robert Guralnick, David Bloom, John Wieczorek, Kyle Braak, Javier Otegui, Laura Russell, and Peter Desmet. The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PLoS ONE*, 2014.
4. 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.
5. Tanu Malik and Ian Foster. Addressing data access needs of the long-tail distribution of geoscientists. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 5348–5351. IEEE, 2012.
6. Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.
7. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.
8. Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.
9. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

10. Marcos Zárate, Pablo Rosales, Pablo Fillottrani, Claudio Delrieux, and Mirtha Lewis. Oceanographic data management: Towards the publishing of pampa azul oceanographic campaigns as linked data. In *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW 2018)*, 2018.
11. Marcos Zárate, Pablo Rosales, Germán Braun, Mirtha Lewis, Pablo Rubén Fillottrani, and Claudio Delrieux. Oceangraph: Some initial steps toward a oceanographic knowledge graph. In Boris Villazón-Terrazas and Yusniel Hidalgo-Delgado, editors, *Knowledge Graphs and Semantic Web*, pages 33–40, Cham, 2019. Springer International Publishing.
12. Marcos Zárate, Carlos Buckle, Renato Mazzanti, Mirtha Lewis, Pablo Fillottrani, and Claudio Delrieux. Harmonizing big data with a knowledge graph: Oceangraph kg uses case. In Enzo Rucci, Marcelo Naiouf, Franco Chichizola, and Laura De Giusti, editors, *Cloud Computing, Big Data & Emerging Topics*, pages 81–92, Cham, 2020. Springer International Publishing.
13. Marcos Zárate, Germán Braun, Mirtha Lewis, and Pablo Fillottrani. Observational/hydrographic data of the south atlantic ocean published as lod. *Semantic Web*, 13(2):133–145, 2022.
14. Derrick Snowden, Vardis M Tsontos, Nils Olav Handegard, Marcos Zarate, Kevin O'Brien, Kenneth S Casey, Neville Smith, Helge Sagen, Kathleen Bailey, Mirtha N Lewis, et al. Data interoperability between elements of the global ocean observing system. *Frontiers in Marine Science*, 6:442, 2019.
15. M. Zárate, G. Braun, and P. Fillottrani. Adding biodiversity datasets from argentinian patagonia to the web of data. *CEUR Workshop Proceedings*, 1963, 2017.
16. Marcos Zárate, Paula F Zermoglio, John Wieczorek, Anabela Plos, and Renato Mazzanti. Linked open biodiversity data (lobd): A semantic application for integrating biodiversity information. *Biodiversity Information Science and Standards*, 4:e58975, 2020.
17. Marcos Zárate, Germán Braun, Pablo Fillottrani, Claudio Delrieux, and Mirtha Lewis. Bige-onto: an ontology-based system for managing biodiversity and biogeography data. *Applied Ontology*, 15(4):411–437, 2020.
18. Sparql 1.1 overview. w3c recommendation 21 march 2013. https://www.w3.org/TR/sparql11-overview/, 2013. [Online; accessed 16-Jul-2022].
19. W3C Owl Working Group et al. Owl 2 web ontology language document overview. *http://www. w3. org/TR/owl2-overview/*, 2009.
20. Mariano Fernández-López, María Poveda-Villalón, Mari Carmen Suárez-Figueroa, and Asunción Gómez-Pérez. Why are ontologies not reused across the same domain? *J. Web Semant.*, 57, 2019.
21. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI global, 2011.
22. Ali Hasnain and Dietrich Rebholz-Schuhmann. Assessing fair data principles against the 5-star open data principles. In *European Semantic Web Conference*, pages 469–477. Springer, 2018.