Visual Analytics for Linked Open Data in Marine Sciences

Gustavo Nuñez[1,*], Carlos Buckle[1][0000−0003−0722−0949], and Marcos Zárate[2][0000−0001−8851−8602]

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

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

Abstract The purpose of data exploration and data visualization (DV) is to offer ways of perceiving and manipulating information, as well as extracting and inferring knowledge. In this short paper we present advances in visual representations and intuitive interaction techniques based on artificial intelligence. This contributes significantly to the exploration and understanding of information related to marine sciences represented by ontologies and linked data. This preliminary research allows scientists and non-expert users to analyze sets with information related to oceanography, meteorology and environmental parameters to promote scientific knowledge and productive innovation in the South Atlantic ocean using Linked Open Data (LOD).

Keywords: Data Visualization · Linked Open Data · Marine Science.

1 Introduction

The purpose of DV is to offer forms of information perception and manipulation, as well as knowledge extraction and inference [1,2]. DV provides to users an intuitive way to explore content, identify interesting patterns, infer correlations and causalities, and supports meaning-construction activities. One of the promising approaches to address the problems associated with integration and graphing is to store in a structured way and reproduce the data sets in graphs. The Semantic Web (SW) [3] offers solutions to these needs by enabling LOD [4] where data objects are uniquely identified and the relationships between them are explicitly defined. LOD is a powerful and compelling approach to disseminating and consuming scientific data from various disciplines [5,6,7,8]. It involves publishing, sharing, and connecting data on the Web and offers different methods of aggregation and interoperability.

* Corresponding author.

In recent years, this way of publishing data has been adopted in a large number of LOD disciplines [9], this has made the visualization and exploration of information a crucial task for most of the LOD consumers. Data scientists, domain experts, and citizens want to use intuitive and visual, rather than programmatic, ways to interact with these resources. In the domain of marine sciences, the visualization of data from disciplines such as Oceanography, Meteorology and Biodiversity face great challenges, since there is an exponential increase in their volume due to the growth of technology and the multiplicity of platforms of remote sensing and the demand for knowledge to contribute globally to climate change models [10]. In addition, there is a great diversity in the type of records that must be displayed properly, the physical chemical, geological, meteorological and biological values must be integrated and the analysis/information products must be based on all of them so that the user can make a correct interpretation [11].

The remainder of this short paper is structured as follows: Section 2 presents different initiatives based on LOD for marine sciences. Section 3 presents a proof concept platform developed to visualize information of marine sciences in South Atlantic. Finally, in section 4, we present some conclusions related to previous experiences and planning for future works.

2 Background and Related Work

A large number of LD visualization tools have been introduced in recent years, most of them originating from academia. DV tools in linked data provide graphical representations of a data set or parts of it, with the aim of facilitating its analysis and generating insights from complex, interlinked information on a geotemporal space. Techniques may vary depending on the domain, the type of record, the task the user is attempting to perform, as well as the user's skills.

There are several initiatives carried out in the Argentine context to publish marine science data such as LD, among them we can highlight: [12] which presents the publication of metadata from oceanographic campaigns as LD. OceanGraph [13] defines an oceanographic knowledge graph prototype to manage information from expeditions, scientific publications and environmental variables, while in [14] OceanGraph exploitation is proposed with concrete examples of potential uses by specialists.

At the international level there are also initiatives for the publication of marine science data such as LD, among the main ones we can mention GeoLink [15], a project funded by the EarthCube initiative, which has taken advantage of the principles of LOD to create a database, which allows users to consult and reason in some of the most outstanding geoscience repositories in the United States. The GeoLink dataset includes such diverse information as port calls made by oceanographic cruises, metadata from physical samples, funding of research projects and personnel, and authorship of technical reports. The data has been published in accordance with best practices for LOD [16] and are publicly available through a SPARQL endpoint that currently contains more than 45 million RDF triples.

3 LD Visualization in Marine Science

In the context of marine science, visual exploration is a promising approach for exploring and analyzing data and better understanding the dynamics of complex ocean processes, although publishing data as LD has a number of success stories [15,17], visualization continues to be a problem because it is a particular task that differs from classic DV, mainly due to the characteristics of LD. The use of common vocabularies (cross domains) for the description of the records or the use of typified properties to capture relationships between resources within a set or between different sets, differs from traditional forms of visualization which are unable to capture the complex possible relationships. For the tests described below, we use public information on marine species and environmental variables captured in the South Atlantic through a SPARQL endpoint whose URL is `http://linkeddata.cenpat-conicet.gob.ar/snorql/`. The methodology used for the creation and publication is detailed in [17]

3.1 Case study

Our focus is on the Web-based front-end consisting of querying and visualization tools. We have developed a proof of concept for interactive visualization of oceanographic, environmental and marine biodiversity information. The platform allows the representation and visualization of interactive maps with trajectories of oceanographic vessels, as well as the retrieval of graph schemes of the relationship between environmental variables and species. For which a selection of open source instruments compatible with the visualization of specific types of information was carried out, for example geospatial data, species distribution, traceability and records related to the environment. Figure 1 shows 2 visualizations used to interpret information on a specific species, in this case Mirounga Leonina (Southern elephant seal). Map shows the information of trips made by several individuals during their feeding trips in the sea, additionally overlapping layers with environmental and special information. The other visualization shows bibliographic information associated with the species.

The application is built with the Shiny framework[3] for the R programming language. Access to endpoints is done through the SPARQL package[4]. The application layout is produced with the flexdashboard package[5], and the maps use Leaflet.js, Highcharts, and ggplot2, all accessed through their corresponding R packages.

To see details of the implementation and the source code, see the following link.

[3] `https://shiny.rstudio.com/`
[4] `https://cran.r-project.org/web/packages/SPARQL/SPARQL.pdf`
[5] `http://rstudio.github.io/flexdashboard/index.html`

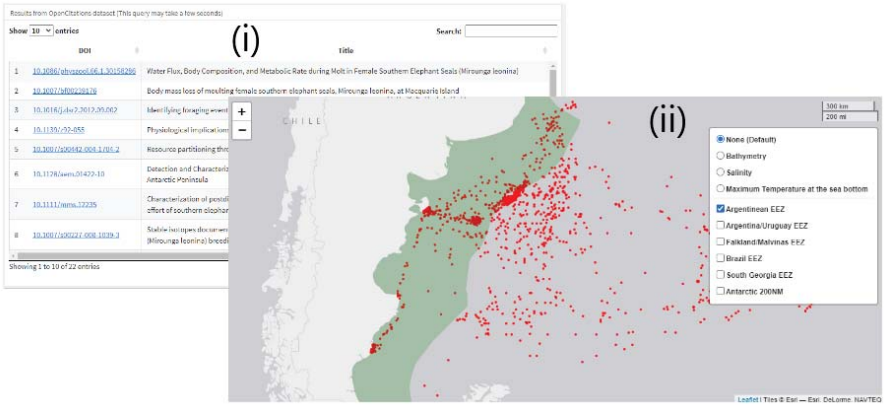

Fig. 1. Visualizations used to relate: (i) marine species with bibliographic information (ii) geo-spatial information of species with environmental variables and marine regions.

4 Conclusions and Future Work

From previous experiences, we can conclude that it is necessary to develop systems capable of visually managing information for comprehensive and secondary use, both from the participating groups and from external users who require information. The results of this preliminary research constitute a substantial contribution, not only for marine sciences, but also as a methodological contribution to scientific visualizations using LD.

As future work, we propose the need to formalize the proof of concept, for this it is necessary to delve into the following aspects: a) Study and research of scientific visualizations typical of marine sciences. b) Develop an online data visualization platform based on prediction models to provide visual analytic facilities and allow interactive queries and analysis of different layers of information. c) Expand the platform or scale the results to other marine spaces, in particular to the Priority Geographic Areas (AGP) of the Pampa Azul initiative[6].

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. Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Communications of the ACM*, 55(4):45–54, 2012.

[6] https://www.pampazul.gob.ar/areas-prioritarias/

2. Stratos Idreos, Olga Papaemmanouil, and Surajit Chaudhuri. Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 277–281, 2015.
3. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.
4. 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.
5. Richard K Lomotey and Ralph Deters. Terms extraction from unstructured data silos. In *System of Systems Engineering (SoSE), 2013 8th International Conference on*, pages 19–24. IEEE, 2013.
6. Syed Ahmad Chan Bukhari, Mate Levente Nagy, Paolo Ciccarese, Michael Krauthammer, and Christopher JO Baker. icyrus: A semantic framework for biomedical image discovery. In *SWAT4LS*, pages 13–22, 2015.
7. Syed Ahmad Chan Bukhari. *Semantic enrichment and similarity approximation for biomedical sequence images*. PhD thesis, University of New Brunswick (Canada), 2017.
8. Roderic D.M. Page. Ozymandias: a biodiversity knowledge graph. *PeerJ*, 7:e6739, April 2019.
9. The open linked data cloud. https://lod-cloud.net/, 2021. [Online; accessed 3-May-2021].
10. Tanu Malik and Ian Foster. Addressing data access needs of the long-tail distribution of geoscientists. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 5348–5351. IEEE, 2012.
11. Alex Hardisty and Dave Roberts. A decadal view of biodiversity informatics: challenges and priorities. *BMC ecology*, 13(1):16, 2013.
12. 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.
13. 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.
14. 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.
15. Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. The geolink knowledge graph. *Big Earth Data*, 2018.
16. Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman. Five stars of Linked Data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.
17. 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.