






Harmonizing Big Data with a Knowledge Graph: *OceanGraph KG* Uses Case

Marcos Zárate^{1,2}(✉) , Carlos Buckle² , Renato Mazzanti^{2,3} ,
Mirtha Lewis¹ , Pablo Fillottrani^{4,5} , and Claudio Delrieux⁶ 

¹ Centre for the Study of Marine Systems,
Patagonian National Research Centre (CENPAT-CONICET),
Puerto Madryn, Argentina

{zarate,mirtha}@cenpat-conicet.gob.ar

² Laboratorio de Investigaciones en Informática (LINVI) - Facultad de Ingeniería,
Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB),
Puerto Madryn, Argentina

³ Unidad de Gestión de la Información, (UGI-CENPAT), Puerto Madryn, Argentina
renato@cenpat-conicet.gob.ar

⁴ Computer Science and Engineering Department,
Universidad Nacional del Sur, (DCIC-UNS), Bahía Blanca, Argentina
prf@cs.uns.edu.ar

⁵ Comisión de Investigaciones Científicas, Provincia de Buenos Aires (CICPBA),
Buenos Aires, Argentina

⁶ Electric and Computer Engineering Department,
Universidad Nacional del Sur (DIEC-UNS), Bahía Blanca, Argentina
cad@uns.edu.ar

Abstract. In this paper we introduce recent efforts carried out by the *OceanGraph KG* project to integrate semi-structured or unstructured content. We present some of the practical applications of *OceanGraph* through use cases, and finally summarize the lessons learned during the development process.

Keywords: Big data integration · Knowledge graph · Linked open data · OceanGraph KG

1 Introduction and Motivation

The management of data generated in several disciplines, including Oceanography and Meteorology, is currently facing great challenges. Among other facts, this is triggered by the recent exponential increase in its volume and diversity of sources, due to the growth of technology and advances in remote ocean observatories [1]. In addition, there is a great diversity in data types that must be handled together. This includes physicochemical, geological, meteorological and biological data, which must be integrated, and the analysis/information products for scientific, governmental, and productive purposes must be based on

integrating all of them to be meaningful [2]. Taking into account the definition of Big Data (BD) [3], both ocean observation and weather data fit within the “5V” characterization of BD (volume, velocity, value, veracity, and variety). Therefore, data management in this context can be considered as a typical Big Data case [4]. In scientific activities, this situation presents both challenges and opportunities regarding the access and integration of data they need to conduct novel research activities that may trigger new discoveries enabled by the integration of multidisciplinary information sources [5, 6]. In the context of the *Horizon 2020 program (H2020)*¹ of the European Union, and at the National level in the strategic plan *Argentina Innovadora 2020*, established by the Ministry of Science, Technology and Innovation (MINCyT) of Argentina, BD and data science are considered fundamental disciplines to address the complexity and scope of the issues that require an interdisciplinary approach and a broad projection in the use of information. In the research activities focused on the South Atlantic, data collection campaigns are scarce, and an adequate information management system is not readily available. Therefore, it is necessary to develop systems capable of managing data integration and delivery, both for the direct and indirect use by the participating research groups and institutions, and for external users that require information (*f.e.*, governmental, third parties, etc.).

One of the advantageous features of BD is its ability to manage information in schema-free formats that are both agnostic with respect to technological aspects, and that allow further schema-evolution that will be typically be the case in Natural Sciences. This allows the use of practical internal representations that facilitate specific purposes, for instance the management of datasets in graph form. The Semantic Web (SW) [7] provides solutions to these needs by enabling the Linked Data (LD) Web [8] where data objects are uniquely identified and the relationships between them are defined explicitly. LD is a powerful and compelling approach to store, disseminate and consume scientific data from various disciplines [6, 9, 10]. LD enables the publication, exchange and connection of data on the Web and offers a new way of integration and interoperability. Recently the term *knowledge graph* (KG) emerged [11], which has been used in research and business, generally in close association with SW technologies, LD, large-scale data analysis and cloud computing. The popularity of KGs is related to the launch of Google Knowledge Graph in 2012², and through the introduction of other large databases by major technology companies, such as Yahoo, Microsoft, AirBnB and Facebook, which have created their own KGs to enhance semantic searches [12]. Not only in the industry there are successful uses of KGs, in the oceanographic domain and in the Life sciences in general there is a growing recognition of the advantages of SW technologies [13–18].

Related to these problems, two previous works were developed for the creation of an Oceanographic linked dataset, both were developed jointly with the *Centro de investigación y transferencia Golfo San Jorge*, (CIT-GSJ-CONICET):

¹ <https://eshorizonte2020.es/>.

² <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>.

the proposal of publication of oceanographic campaign metadata [19], and the definition of initial steps for the development of an oceanographic KG called *OceanGraph KG* [20]. Based on the experience gained in this previous work, a series of recommendations related to interoperability and information integration of The Integrated Ocean Observing System (IOOS) [21] was proposed.

This paper describes in a general way the *OceanGraph KG* and its recent efforts focused on the integration of heterogeneous oceanographic and meteorological data. In Sect. 2 we present the underlying idea of *OceanGraph KG* and its main features. In Sect. 3 we discuss its usefulness through case studies. Finally, in Sect. 4 lessons learned and future guidelines are presented.

2 OceanGraph KG Overview

The first version developed to integrate heterogeneous data taking advantage of a KG was described in [20]. *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 *National Marine Data System (NMDS)*³, datasets they collected (from *Global Biodiversity Information Facility (GBIF)*⁴ and *Ocean Biogeographic Information System (OBIS)*⁵, and papers written by themselves (from Springer Nature SciGraph)⁶. 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).

2.1 Ontologies and Vocabularies Used

To ensure that our data will be available to multiple scientific communities, the resource description should adopt well-known standards. Next, we will describe the main resources related to the oceanographic domain and we will see the selected standards to model information on agents and organizations. Different data providers use their own ontologies and reuse existing ones.

- **National Environmental Research Council's (NERC) Vocabulary Server (NVS)** [14] provides access to standardized lists of terms which are used to facilitate data mark-up, interoperability and discovery in the marine science domain. NVS is published as Linked Data on the web using the data model of the Simple Knowledge Organization System (SKOS)⁷.

³ <http://www.datosdelmar.mincyt.gob.ar/index.php>.

⁴ <https://www.gbif.org/>.

⁵ <http://www.iobis.org/>.

⁶ <https://www.springernature.com/gp/researchers/scigraph>.

⁷ <https://www.w3.org/2004/02/skos/>.

- **GeoSPARQL** [22] defines an ontology that supports geospatial semantics, developed by the Open Geospatial Consortium (OGC)⁸. The definition of this ontology (based on well-known OGC standards) is intended to provide a basis for the standardized exchange of RDF geospatial data that can offer query capabilities and qualitative spatial reasoning using the W3C standard SPARQL [23].
- **Darwin Core Standard** [24] provides a stable, direct and flexible structure for compiling and sharing biodiversity data from different sources. *OceanGraph*, uses it to describe properties and concepts related to occurrences of marine species.
- **Geolink** [15] dataset includes diverse information, such as port stops made by oceanographic cruises, physical sample metadata, funding for research projects and staff. This dataset is based on an ontological design pattern (ODP). This ODP it is generic enough to adapt it to the modeling needs established by *OceanGraph*.
- **BiGe-Onto** [25] is an ontology designed to manage Biodiversity and Marine Biogeography data. *BiGe-Onto* uses the idea of *occurrence* (the observation of a species in a place at a given time), since the censuses are observations of SES at a specific time and place, we consider that *BiGe-Onto* fits to nature of our data. *BiGe-Onto* also reuses different appropriate vocabularies to represent information from these domains. In particular, Darwin Core (DwC) [24] is the most important thereof, and reuses several classes that will be considered here: *Occurrence*, *Event*, *Taxon* and *Organism*. *BiGe-Onto* also reuses *foaf:Person*, *void:Dataset* and *dcterms:Location*. Our ontology models occurrences that are related to other concepts through the following relationships.
 - *bigeonto:associated*. Each of the occurrences are described according to the existence of an organism, which was observed at a specific place and time. The organism and the taxon are related through *bigeonto:belongsTo* property.
 - *bigeonto:has.event*. The occurrence has a location (since they are species observations) and they are given by the relation *bigeonto:has_location*, which belongs to a specific environment *bigeonto:characterizes*. The Relations Ontology (RO)⁹ defines the relationships between *bigeonto:Environment* and the classes of the Environment Ontology (EnvO) [26].
 - *dwciri:recordedBy*. This property enables non-literal ranges in comparison to its analog *dwc:recordedBy*, so it allows to relate URIs that describe people, groups or organizations involved in the occurrence, *e.g.* relate a person to their ORCID.
 - *dwciri:inDataset*. Allows the occurrences to be related to the data set to which they belong.
- **SSN/SOSA** [27] To describe the sensors and their oceanographic observations, we use the Semantic Sensor Network (SSN) ontology, and especially the Sensor, Observation, Sample and Actuator (SOSA) ontology that describes the elemental classes and properties, for example (depth, temperature, salinity, etc.).

⁸ <http://www.opengeospatial.org/>.

⁹ <https://github.com/oborel/obo-relations>.

Both vocabularies are suitable for a variety of applications, like large-scale scientific monitoring, satellite imagery, among others. The SSN ontology is an OWL vocabulary developed by the W3C, in collaboration with the Open Geospatial Consortium (OGC), so its adoption guarantees its reuse in many other applications.

2.2 Cross-linking

A challenge, in order to improve the discovery of information, is to generate links between the different URIs of the KG. The interlinking of *OceanGraph* data sets was carried out semi-automatically. It is common for people who participated in an oceanographic campaign, after it, to publish their results in scientific journals. Even more complex is the case of a person who publishes a *datapaper* (scientific paper that describes data), this is made up of the publication itself, plus the primary data that supports it in OBIS or GBIF. *OceanGraph* allows people or species to be linked in different repositories, thus ensuring semantic interoperability between data sets. To generate the links we use the SILK framework¹⁰, which uses the declarative language Silk-LSL (Link Specification Language) with which the user can establish the type of RDF links that must be discovered between the different data sets and the conditions that must be met, *e.g.* to relate researchers who obtained data from a campaign with the results published in OBIS or GBIF, the *Levenshtein distance* is used to disambiguate entities by calculating the similarity between them.

This operator receives two inputs: `dwc:recordedBy`¹¹ and `foaf:name`, if there is enough match that the people are the same, SILK generates the link between them using the axiom `owl:sameAs`. Figure 1 shows the relationships used to integrate *OceanGraph* datasets.

2.3 Availability

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 GraphDB¹² 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:

¹⁰ <http://silkframework.org/>.

¹¹ <https://terms.tdwg.org/wiki/dwc:recordedBy>.

¹² <http://graphdb.ontotext.com/>.

ocean.user). *OceanGraph KG* is also available for download in [28] under CC BY 4.0 license. Table 1 summarizes the main links to explore the knowledge graph in various ways.

Table 1. Main features of *OceanGraph KG*.

Feature	URL
Repository name	OceanGraph (user: oceangraph password: ocean.user)
Repository URL	http://web.cenpat-conicet.gob.ar:7200/login
SPARQL endpoint	http://web.cenpat-conicet.gob.ar:7200/OceanGraph
Visual SPARQL endpoint	http://web.cenpat-conicet.gob.ar:7200/sparql
Class hierarchy	http://web.cenpat-conicet.gob.ar:7200/sparql
Vocabularies	19
No. classes	23
No. properties	50
No. triplet	4.6 M

3 Big Data Use-Cases

As a result of the process described in the previous sections, a set of nodes and links were created to connect references from the input data to entities and relationships within the KG. We extended this generic approach to integrate different functionality modes that are typical in BD contexts.

3.1 Complementing Information with SN SciGraph

As the development and adoption of novel research devices is growing exponentially, it's getting harder to track all the documents related to a given scientific subject. SciGraph dataset integrates data sources from Springer Nature. SciGraph collects information about research landscape: research projects, publications, conferences, funding agencies and others. This dataset [29] includes around 35 million records and is refreshed on a monthly basis.

It is often necessary to connect researchers or other stakeholders that contribute to the same subject. This is specifically the case in the oceanographic domain, in which is required to determine researchers who are part of an oceanographic campaign, and connect their subject with other researchers from another part of the world who are working on the same subjects. In the particular case study of this paper, the research subject is physical oceanography.

We will explore the instances of the *sg:Subject* class and their related subjects using the *core#narrower* property. As can be seen in Fig. 2, there are five subjects directly related to physical oceanography (*ocean science*, *marine biology*, *climate sciences*, etc.)

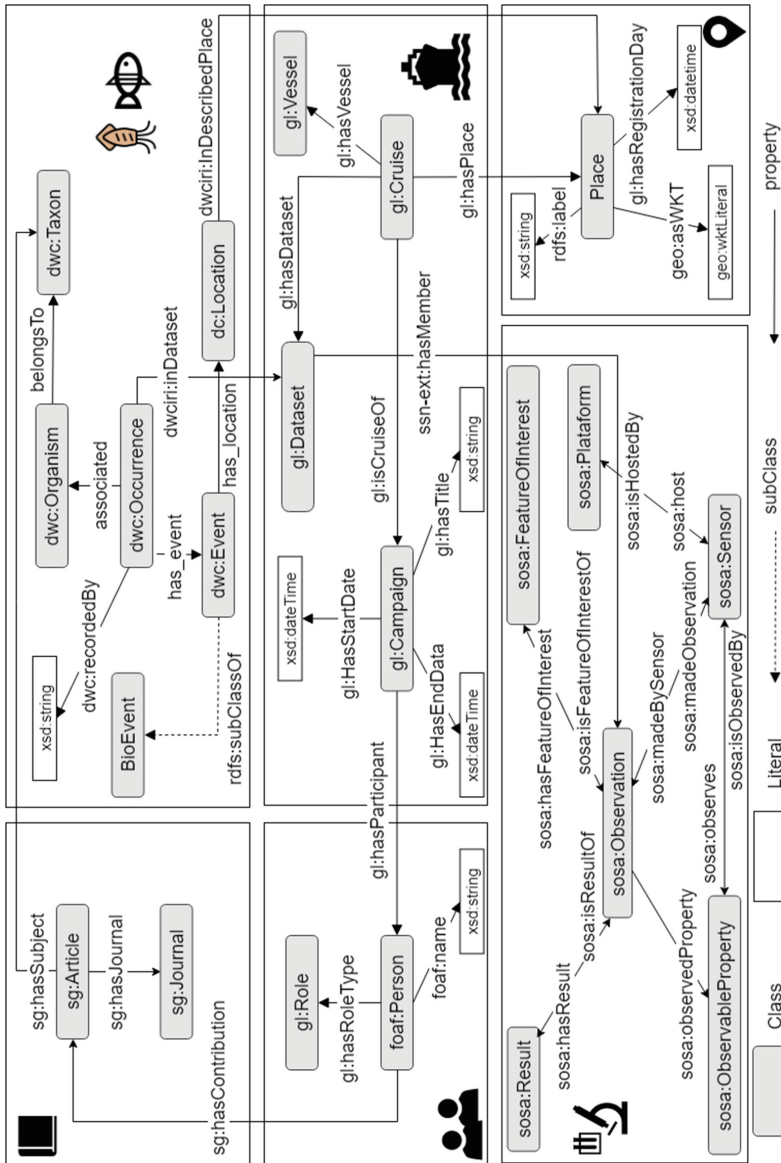


Fig. 1. Conceptual diagram of *OceanGraph KG*. For simplicity, only the main object properties are shown, which allow relationships between the classes of each data set to be established.

Visual graph ❗

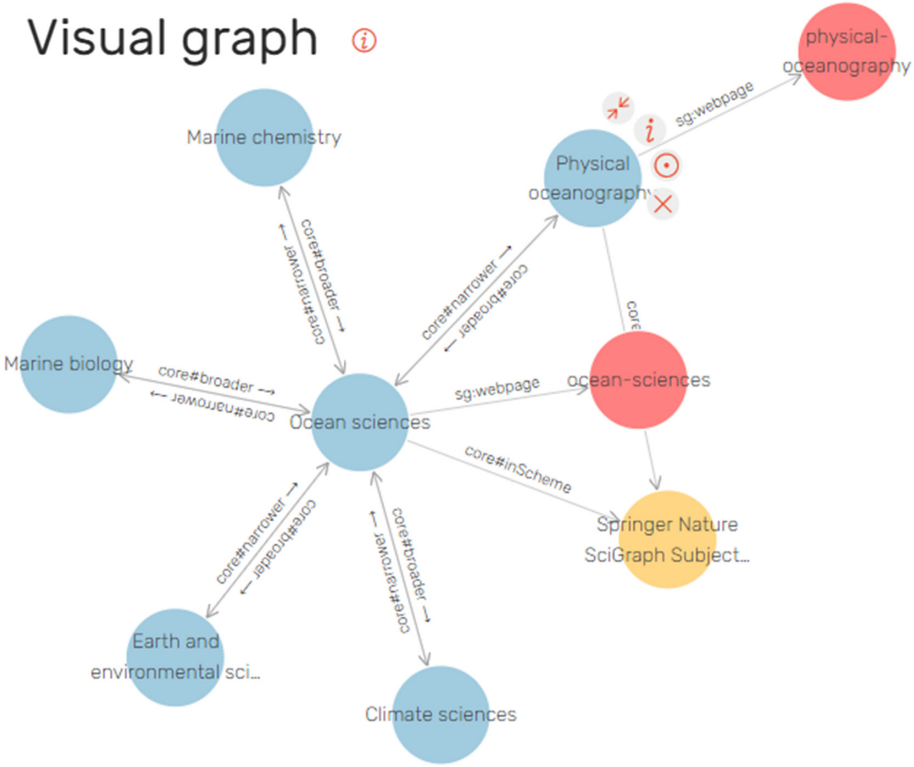


Fig. 2. Exploring terms related to the concept *physical oceanography* using the GraphDB visual interface.

3.2 Macroecological Analyzes

A very common requirement of macroecological analyzes, particularly those that consider the environmental drivers of species distributions, is to match occurrences of species' with environmental variables, and how distributions are expected to shift as the climate changes. This case study shows an example of how KG information can be exploited using the relationships between occurrences with environmental variables, for the example we will use the body of water temperature as a study variable. In particular, we need to associate the following variables: (i) the occurrence of a species under study, (ii) the region of interest, (in our case Golfo Nuevo), (iii) a specific time frame and (iv) the measurements of the water body temperature.

The first step is to define the region under study, to later and then recover the occurrences of the chosen species in a specific time frame. To handle temporal concepts, we use Time Ontology [30]. Since NERC provides URIs for each of the variables that we need to analyze, we only need to search for the URI of the

body water temperature, which is defined as: [SDN:P01::TEMPCU01](#). Table 2 shows an RDF fragment that includes the concepts involved in performing the analysis.

Table 2. RDF serialization of the concepts involved in macroecological analysis.

```

bigeonto:ExtendedMeasurementOrFact a owl:Class.

bigeonto:mesasurement1 rdf:type bigeonto:ExtendedMeasurementOrFact;
rdfs:label "Medicion de temperatura de la columna de agua";
dwc:MeasurementTypeID http://vocab.nerc.ac.uk/collection/P01/current/TEMPCU01/;
dwc:MeasurementValue 6^^xsd:integer;
dwc:MeasurementUnitID http://vocab.nerc.ac.uk/collection/P06/current/UPAA/;
bigeonto:has_event bigeonto:bioevent/urncatalogcenpat-conicet-peces-p-331
bigeonto:has_occurrence bigeonto:occurrence/urncatalogcenpat-conicet-peces-p-331.

bigeonto:occurrence/urncatalogcenpat-conicet-peces-p-331
rdf:type dwc:Occurrence
dwciri:recordedBy http://www.cenpat-conicet.gob.ar/resource/person/unknown;
dwc:basisOfRecord "HumanObservation"^^xsd:string;
dwc:catalogNumber "CNP-P-331"^^xsd:string;
dwc:collectionCode "CNP-PECES"^^xsd:string .

bigeonto:bioevent/urncatalogcenpat-conicet-peces-p-331
rdf:type bigeonto:BioEvent;
dwc:eventDate "08/02/1983"^^xsd:date;
bigeonto:has_location bigeonto:location/urncatalogcenpat-conicet-peces-p-331

```

In Listing 1.1, you can see the query that we implemented using SPARQL, it associates the occurrence of *Merluccius hubbsi* (a fish species of specific scientific and productive interest) with the temperature in a particular region. To do this, we define *Golfo Nuevo*, as an instance of (*geo:Polygon*), then look for observations of *Merluccius hubbsi*, which has its location associated and are instances of the class (*geo:point*). One of the advantages of adopting GeoSPARQL is that we can perform spatial operations, *e.g.* to determine if a point is contained within a polygon, for this we use the provided function (*geof:sfWithin*). As a last step, we must obtain the temperature (also georeferenced) and define it by NERC as *TEMPCU01*. To execute the query in GraphDB, see the following link¹³. This specific example shows how our proposed data integration effort around KGs, bridges the gap between the sometimes isolated existing data collection initiatives worldwide, and a centralized and uniform data access that may be automated. A standardization like the provided by our proposal further enables the next and more fruitful BD stages, including massive automated data analysis, online real-time actionable dashboards, and visual analytics.

¹³ <http://web.cenpat-conicet.gob.ar:7200/sparql?savedQueryName=OG-Q001>.

Listing 1.1. Query required to associate observational occurrences of a particular species within a given geographic region and with specific environmental conditions.

```

PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX bigeonto: <http://www.w3id.org/cenpat-gilia/bigeonto/>
PREFIX gl: <http://schema.geolink.org/1.0/base/main#>
PREFIX geosparql: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX nerc: <http://vocab.nerc.ac.uk/collection/P01/current/>

SELECT ?occ ?measurement ?PointWKT
WHERE {
  ?occ a dwc:Occurrence.
  ?occ bigeonto:associated ?organism.
  ?organism bigeonto:belongsTo ?taxon.
  ?taxon dwc:scientificName ?sciname.
  ?occ bigeonto:memberOf ?dataset.
  ?dataset gl:hasMeasurementType ?measurement.
  ?occ bigeonto:has_event ?event.
  ?event dwc:eventDate ?date.
  ?event bigeonto:has_location ?location.
  ?location geosparql:hasGeometry ?point.
  ?point geosparql:asWKT ?PointWKT.
  bigeonto:polygon/golfo-san-matias-polygon geosparql:asWKT ?PWKT.
  FILTER (geof:sfWithin(?PointWKT, ?PWKT))
  FILTER(regex(str(?measurement), "TEMP" ) )
  FILTER regex(STR(?sciname), "Merluccius hubbsi")
  FILTER (?date >= xsd:date("date") && ?date < xsd:date("date"))
}

```

4 Conclusion

Based on the results of this experience, KGs proved to be powerful and flexible enough to integrate diverse data sets. However, the integration process required to correctly map input data into a KG can be exhausting, since automated techniques have so far been unable to fully understand the semantics of input data. Through the *OceanGraph* development process, we learned a few lessons on how LD can contribute to addressing important BD challenges, especially within the area of oceanographic data.

First, the amount of linked datasets grows every year and is interrelated over a growing entanglement of scientific information. This presents new challenges, which require considering scalability and performance as crucial aspects for any future facility [31]. Around this issue is where LD needs to incorporate BD techniques and methodologies, specifically in the data management aspects.

Second, from the BD perspective, it is also a priority to start incorporating linked data results. Currently only a few large companies are able to take advantage of BD [32], which is unfortunate since individual scientists, small research groups, nongovernmental agencies, and other stakeholders that are engaged in potentially relevant activities are in a disadvantageous situation among the large commercial interest groups. In this foreseeable scenario, some questions that arose in other contexts begin to be visible. Among others we can mention [33]: *How can particular users delve into BD in a fruitful manner? Having found useful data, How to make it understandable to laypersons with little or no prior data science knowledge? How to handle data in a way that grants no privacy or licensing breaches? How can data generated from different cultures and over different languages (or even charsets) be rendered useful? What standards for data and metadata are necessary? How to link data from different repositories?*

What governance standards should be supported or even enforced to grant privacy, traceability, auditing, and other technical, ethical and legal features that systems like this must implement?.

BD is doomed to arrive into the realms of worldwide scientific enterprises, but its value will increase, and all the community will be able to take advantage of it, only when it becomes transparent and often usable by the largest number of users [34]. From this perspective, it is necessary to consider that BD, at least in the context of scientific enterprises, requires multi- and interdisciplinary integration, and, within such a decentralized scenario, the new challenges are also associated with meaning.

Acknowledgments. This work is partially funded by project *Linked Open Data Platform for Management and Visualization of Primary Data in Marine Science*. Supported by Secretariat of Science and Technology of the National University of Patagonia San Juan Bosco (UNPSJB). Some of the data used were provided by the *Golfo San Jorge Research and Transfer Center* (CIT-GSJ-CONICET).

References

1. Malik, T., Foster, I.: Addressing data access needs of the long-tail distribution of geoscientists. In: 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 5348–5351. IEEE (2012)
2. Hardisty, A., Roberts, D.: A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecol.* **13**(1), 16 (2013). <https://doi.org/10.1186/1472-6785-13-16>
3. Beyer, M.A., Laney, D.: The importance of “big data”: a definition, pp. 2014–2018. Gartner, Stamford, CT (2012)
4. Liu, Y., Qiu, M., Liu, C., Guo, Z.: Big data in ocean observation: opportunities and challenges. In: Wang, Y., Yu, G., Zhang, Y., Han, Z., Wang, G. (eds.) *BigCom 2016*. LNCS, vol. 9784, pp. 212–222. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42553-5_18
5. Campbell, P.: Data’s shameful neglect. *Nature* **461**(7261), 145 (2009)
6. Lomotey, R.K., Deters, R.: Terms extraction from unstructured data silos. In: 2013 8th International Conference on System of Systems Engineering (SoSE), pp. 19–24. IEEE (2013)
7. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. *Sci. Am.* **284**(5), 28–37 (2001)
8. Bizer, C., Heath, T., Berners-Lee, T.: Linked data: the story so far. In: *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227. IGI Global (2011)
9. Bukhari, S.A.C., Nagy, M.L., Ciccarese, P., Krauthammer, M., Baker, C.J.: iCyrus: a semantic framework for biomedical image discovery. In: *SWAT4LS*, pp. 13–22 (2015)
10. Bukhari, S.A.C.: Semantic enrichment and similarity approximation for biomedical sequence images. Ph.D. thesis, University of New Brunswick (Canada) (2017)
11. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. 48 (2016)
12. Ceravolo, P., et al.: Big data semantics. *J. Data Semant.* **7**(2), 65–85 (2018). <https://doi.org/10.1007/s13740-018-0086-2>

13. Leadbetter, A., Arko, R., Chandler, C., Shepherd, A., Lowry, R.: Linked data an oceanographic perspective. *J. Ocean Technol.* **8**(3), 7–12 (2013)
14. Leadbetter, A., Lowry, R., Clements, D.O.: The NERC vocabulary server: version 2.0. In: *Geophysical Research Abstracts*, vol. 14 (2012)
15. Krisnadhi, A., et al.: An ontology pattern for oceanographic cruises: towards an oceanographer’s dream of integrated knowledge discovery (2014)
16. Cheatham, M., et al.: The GeoLink knowledge graph. *Big Earth Data* **2**(2), 131–143 (2018)
17. Page, R.D.M.: Ozymandias: a biodiversity knowledge graph. *PeerJ* **7**, e6739 (2019)
18. Springer Nature SciGraph (2018). <http://www.springernature.com/gp/researchers/scigraph>. Accessed 24 Jan 2019
19. Zárate, M., Rosales, P., Fillotrani, P., Delrieux, C., Lewis, M.: 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)
20. Zárate, M., Rosales, P., Braun, G., Lewis, M., Fillotrani, P.R., Delrieux, C.: *Ocean-Graph*: some initial steps toward a oceanographic knowledge graph. In: Villazón-Terrazas, B., Hidalgo-Delgado, Y. (eds.) *KGSWC 2019. CCIS*, vol. 1029, pp. 33–40. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21395-4_3
21. The Integrated Ocean Observing System (IOOS) (2013). <https://ioos.noaa.gov/>. Accessed 19 July 2019
22. Battle, R., Kolas, D.: Enabling the geospatial semantic web with parliament and GeoSPARQL. *Semant. Web* **3**(4), 355–370 (2012)
23. SPARQL query language for RDF (2008). <https://www.w3.org/TR/rdf-sparql-protocol/>. Accessed 10 Mar 2019
24. Wiczorek, J., et al.: Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* **7**(1), e29715 (2012)
25. Zárate, M., Braun, G., Fillotrani, P.R., Delrieux, C., Lewis, M.: BiGe-Onto: an ontology-based system for managing biodiversity and biogeography data. *Appl. Ontol. J.* (2019, accepted paper)
26. Buttigieg, P.L., Pafilis, E., Lewis, S.E., Schildhauer, M.P., Walls, R.L., Mungall, C.J.: The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semant.* **7**, 57 (2016). <https://doi.org/10.1186/s13326-016-0097-6>
27. W3C: Semantic Sensor Network Ontology (SSN) W3C Recommendation (2017)
28. Zárate, M., Buckle, C., Mazzanti, R., Fillotrani, P., Delrieux, C., Lewis, M.: OceanGraph RDF dataset (2020). <https://doi.org/10.17632/9t5xkt9wwk.1>. Accessed 18 Mar 2019
29. Michele Pasin and FigShare Admin SN SciGraph. Dataset: Persons, April 2019
30. Time Ontology in OWL W3C Recommendation 19 October 2017 (2017). <https://www.w3.org/TR/owl-time/>. Accessed 27 Jan 2020
31. Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: a survey of the state of the art. arXiv preprint [arXiv:1601.08059](https://arxiv.org/abs/1601.08059) (2016)
32. Hernández-Pérez, T.: In the age of the web of data: first open data, then big data. *El profesional de la información (EPI)* **25**(4), 517–525 (2016)
33. Hendler, J.: Broad data: exploring the emerging web of data. *Big Data* **1**(1), 18–20 (2013)
34. Manyika, J.: Big data: the next frontier for innovation, competition, and productivity (2011). http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation. Accessed 29 Jan 2020