Oceanographic Data Management: Towards the Publishing of Pampa Azul Oceanographic Campaigns as Linked Data

Marcos Zárate[1,5], Pablo Rosales[2,5], Pablo Fillottrani[4], Claudio Delrieux[3,5], and Mirtha Lewis[1,2]

[1] Center for the Study of Marine Systems, (CENPAT-CONICET), Argentina.
{zarate,mirtha}@cenpat-conicet.gob.ar
[2] Centro de Investigaciones y Transferencia Golfo San Jorge, (CONICET), Argentina.
prosales@unpata.edu.ar
[3] Electric and Computer Engineering Department, (UNS), Argentina.
cad@uns.edu.ar
[4] Computer Science and Engineering Department, (UNS), Argentina.
prf@cs.uns.edu.ar
[5] Universidad Nacional de la Patagonia San Juan Bosco, (UNPSJB), Argentina.

1 Introduction and Motivation

Pampa Azul is a governmental initiative in Argentina that supports research activities through oceanographic campaigns and promotes interdisciplinary cooperation between institutes of marine research in areas of national jurisdiction. The Argentine Continental shelf houses commercial fisheries, biodiversity, hydrocarbon basins and mineral deposits of great economic and ecologic importance. In particular, the San Jorge gulf[6] has ecologic and geographic features that brought together broad and sustained oceanographic research activities. The gulf is located in the central region of the Patagonian ocean litoral, where oil and fishing industries coexist along with tourism, which gives rise to ongoing and acute environmental risks. For this reason, accurate and frequent oceanic sampling and measurement is of critical social, economic and ecologic importance. Oceanographic campaigns funded by the Pampa Azul initiative are the basis of scientific research at sea, yielding huge amounts of data, highly heterogeneous in types and formats, and scattered across distributed data repositories.

Oceanographic research and efficient management of the collected data often appear to be two widely separated worlds. Data managers consider the careful collection, management and dissemination of research data as essential for the effective use, while researchers consider data management as a merely technical issue, of little relevance for their interests. Consequently, data management is often insufficiently planned, if at all, and receives very low priority and budget. As part of governmental policies towards ocean management, researchers from Argentine scientific institutions are required to disseminate their activities and accomplishments to give greater visibility to the national efforts in this field.

[6] http://www.pampazul.gob.ar/areas-prioritarias/golfo-san-jorge/

For individual researchers, this situation presents a difficult challenge regarding discovery, access, and integration of data which they need to conduct scientific inquiries. As specific cases of these problems, we can mention (i) *data level conflicts* caused by differences that arise in data domains due to multiple possible representations and similar data interpretations, and (ii) *lack of an integrated knowledge infrastructure* which hinders the ability of researchers to analyze potential discovery scenarios if more than one repository is involved. For example, they may be interested in knowing if in a marine region there are CTD data available[7], provided by oceanographic campaigns in the region of interest performed by research vessels not belonging to Pampa Azul.

The Semantic Web (SW) [1] provides possible solutions to these and other problems by enabling the web of Linked Data (LD) [2], which is a methodology for publishing data and metadata in a structured format in a way such that links may be created and exploited between objects. The key enabling components are URIs, HTTP, the Resource Description Framework (RDF), and the SPARQL Protocol and RDF Query Language (SPARQL).

In this short paper we present the initial steps in the creation of a oceanographic linked dataset using information from the oceanographic campaigns of Pampa Azul. To achieve consistency, discoverability and make the datasets readable by machines and humans, we use different controlled vocabularies among them NERC Vocabulary Server[8] together with the geospatial standard for the semantic web GeoSPARQL [3] and the reuse of the ontological design pattern (ODP) for oceanographic cruises [4]. In addition we complement the Pampa Azul information with the oceanographic linked dataset Rolling Deck to Repository (R2R[9]).

2 Creating the RDF Dataset

The publication of the dataset involves different steps that are described below and the architecture of such a process can be consulted in[10].

- **Imput data:** *National Marine Data System* (NMDS[11]) is a web platform that allows publishing datasets of oceanographic campaigns that was sampled in the Argentine sea. These datasets are composed of (i) metadata of the oceanographic campaigns (name of the campaign, vessel, dates, people and institutions involved, geographical coverage among others.) in XML format, and (ii) data recorded by the vessel in its trajectory, which contains the information of the measured variables (pressure, salinity, temperature, depth,

[7] CTD: instrument used to measure the conductivity, temperature, and pressure of seawater.

[8] `https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/P01/`

[9] `http://data.rvdata.us/`

[10] `https://github.com/cenpat/pa-lod/blob/master/images/framework-PA.png` accessed at April 2018

[11] `http://www.datosdelmar.mincyt.gob.ar/index.php`

positions where the variable was sampled among others) and additionally contains information of the equipment that was used to sample (CTD, Termosalinometer, etc.).

- **Data extraction and cleaning:** Metadata and data of campaigns are manually extracted from the NMDS repository and their content are processed using OpenRefine tool[12]. There, the columns are cleaned and converted to standardised data types such as dates, numerical values, etc. and empty columns are removed.
- **URI strategy:** Currently URIs for the resources belonging to oceanographic campaigns follow the pattern:

  `http://data.pa.gob.ar/lod/{type}/{concept}/{ID}`

  The domain only be used for the publication of PAMPA AZUL information and not include the name of any organization, as they may evolve over time. {`type`} can take any of the following values: `resource` for the HTTP URI of a resource, and `page` and `data` for that resource's HTML and RDF documents respectively. {`concept`} gives a hint as to what this resource is about by referring to the class to which that resource belongs, for example, `Cruise, Dataset, Person` etc. {`ID`} for the unique identifiers we use the ones provided in the original datasets, normally identified with a Universally Unique Identifier (UUID).
- **Conversion to RDF:** Data are converted to RDF triples using RDF Refine[13] that allows users to go through a graphical interface describing the RDF scheme skeleton which specifies the subject, predicate and the object of the triples to be generated. The next step in the process is to set up prefixes. Since datasets include localities, locations and research institutes, we set up prefixes for well-known vocabularies such as FOAF, Dublin Core, NERC parameter codes and GeoSPARQL. To see the resulting graph after the conversion see the link[14].
- **RDF storage:** The transformed data have been published, and can to be visualised through GraphDB which is a highly efficient and robust graph database with RDF and GeoSPARQL support. It allows users to explore the hierarchy of RDF classes, relationships among these classes, etc.

3 Use Case: Complementing Pampa Azul Information With R2R

The following use case explores the R2R oceanographic linked dataset using a federated query to retrieve all the trajectories that exist in the R2R dataset and that are within a polygon defined in the `FILTER` clause, this polygon defines the exclusive Argentine economic area, so this query is interesting since it allows to know which cruises traveled that region at some time.

[12] `http://openrefine.org/`

[13] `http://refine.deri.ie/`

[14] `https://github.com/cenpat/pa-lod/blob/master/images/pa-graph.png` accessed at April 2018

```
PREFIX geosparql: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/geosparql/function/>
PREFIX sf: <http://www.opengis.net/ont/sf#>

SELECT ?track
WHERE {
 SERVICE <http://data.rvdata.us/sparql> {
   ?track a sf:LineString.
   ?track geosparql:asWKT ?gWKT
   FILTER (geof:sfWithin(?gWKT, "POLYGON((-63.80859375 -41.45713698292349,
   -47.109375 -41.457136982923494, -47.109375 -50.9151558824997, -63.80859375
   -50.9151558824997, -63.80859375 -41.457136982923494))"^^sf:WktLiteral)) }
}
```

4 Conclusion and Future Work

In this short paper we presented an overview of our initial efforts to create a Linked Oceanographic Dataset, reusing Ontological Design Patterns and specific vocabularies of this domain. In order to test our dataset we extracted metadata in XML format from NMDS. In this initial stage (April 2018) our platform stored 618K RDF triples with a total of ten classes instantiated. Also for the user to exploit the dataset we define SPARQL queries that can be accessed through the link[15]. Finally, the user can visually explore the dataset, accessing the following link to the GraphDB interface[16] (user: `pauser` password: `pauser`). We have shown that RDF can be used to represent metadata about oceanographic campaigns of Pampa Azul initiative in a useful way, but there is still a lot of fruitful work to be done. Particularly as future work, we need to develop information visualization interfaces that allow non-expert users to explore the data. For this we have explored map4rdf[17] a mapping and faceted browsing tool for exploring and visualizing RDF datasets enhanced with geometrical information.

References

1. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The Semantic Web. *Scientific American*, 2001.
2. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, 2009.
3. Robert Battle and Dave Kolas. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, 3(4):355–370, 2012.
4. Adila Krisnadhi, Robert Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Timothy Finin, Pascal Hitzler, Krzysztof Janowicz, Thomas Narock, Lisa Raymond, et al. An ontology pattern for oceanographic cruises: Towards an oceanographer's dream of integrated knowledge discovery. 2014.

[15] `https://github.com/cenpat/pa-lod/tree/master/SPARQL` accessed at April 2018
[16] `http://web.cenpat-conicet.gob.ar:7200/login`
[17] `http://oegdev.dia.fi.upm.es/projects/map4rdf/`