BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e58975
doi: 10.3897/biss.4.58975

OPEN ACCESS

Conference Abstract

Linked Open Biodiversity Data (LOBD): A semantic application for integrating biodiversity information

Marcos Zárate[‡], Paula F Zermoglio[§], John Wieczorek[§], Anabela Plos[|], Renato Mazzanti[¶]

[‡] Center for the Study of Marine Systems (CESIMAR-CENPAT-CONICET), Puerto Madryn, Argentina
[§] VertNet, Bariloche, Argentina
[|] Museo Argentino de Ciencias Naturales "Bernardino Rivadavia" MACN-CONICET, GBIF Argentina, Buenos Aires, Argentina
[¶] CENPAT-CONICET, Puerto Madryn, Chubut, Argentina

Corresponding author: Marcos Zárate (zarate@cenpat-conicet.gob.ar)

Received: 25 Sep 2020 | Published: 28 Sep 2020

Citation: Zárate M, Zermoglio PF, Wieczorek J, Plos A, Mazzanti R (2020) Linked Open Biodiversity Data (LOBD): A semantic application for integrating biodiversity information. Biodiversity Information Science and Standards 4: e58975. https://doi.org/10.3897/biss.4.58975

Abstract

Scientists frequently collect biological and environmental information over years and store it in database systems to answer their own research questions without exposing it in repositories that make it easy to find and retrieve.

While in recent years the community working on biodiversity informatics has made significant strides by creating common shared vocabularies such as the Darwin Core (DwC, Wieczorek et al. 2012) and publishing mechanisms such as the Integrated Publishing Toolkit (IPT, Robertson et al. 2014), integration is largely limited to the aggregation of datasets and full interoperability has still not been achieved. In this context, The Semantic Web (SW) aims to represent information in a way that, in addition to the human-centered display purposes, it can be used autonomously by machines for integration and reuse across applications. From the biodiversity informatics point of view, interoperability and links among data sources would allow integration of information that is otherwise disconnected, enabling scientists to answer broader questions. These

considerations provide strong motivations to formulate a web application considering the semantic interoperability that may provide answers to questions such as the following:

- (Q1) Is it possible to complement taxonomic, bibliographic and environmental information of a particular species without relying on specific Application Programming Interfaces (APIs)?
- (Q2) How to relate occurrences of species with environmental variables within a specific region?
- (Q3) What are the bibliographic references associated with a given species?

With questions such as these in mind, we present the design of a proof-of-concept application: Linked Open Biodiversity Data (LOBD). LOBD uses Linked Data (LD) (Heath and Bizer 2011) to complement species occurrence information previously extracted from GBIF and converted to Resource Description Framework (RDF) (Zárate et al. 2020) with information about the taxa in question from different RDF datasets, such as Wikidata, NCBI Taxonomy, Springer Nature SciGraph and OpenCitation corpus. A simplified view of the architecture is shown in Fig. 1. To achieve semantic interoperability, we use the SPARQL query language, which allows us not to depend on specific APIs to retrieve information.

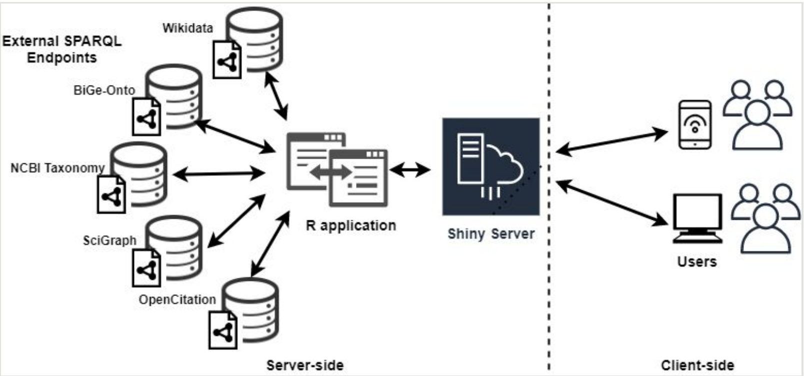

Figure 1.

Simplified view of the LOBD client/server architecture. The client performs a query and request through a web browser. Depending on the request, the server consults the different SPARQL endpoints and returns the results through the Client-side user interface..

The application consists of three modules:

1. **General information**, where the Wikidata endpoint is used to retrieve additional information about the selected species, including links to other databases and information about the species extracted from National Center for Biotechnology Information (NCBI) Taxonomy.
2. **Bibliography**, where all publications related to the species are retrieved and extracted from OpenCitation.
3. **Environment**, where users can plot species on a map and add layers related to marine regions as well as environmental layers (e.g., temperature, salinity, etc).

For the development of the application, we use the Shiny framework for R, access to SPARQL endpoints is done through the SPARQL package, marine regions are obtained from marineregion.org and the environmental layers are extracted from Bio-ORACLE.

The data used for this article were collected by the Center for the Study of Marine Systems at the National Patagonian Sci-Tech Centre (CCT CENPAT-CONICET), and are published and available through the GBIF network.

Linked Data is a powerful tool for scientists, as it allows generating new approaches to biodiversity informatics, which can help to address the data integration challenges. Users would benefit from complementing the current prevalent use of vocabularies that are not ontologically defined (like DwC) for sharing biodiversity data. Although this application is a proof of concept, it shows that with little effort, it is possible to achieve greater interoperability between datasets that were not initially represented as LD.

Keywords

semantic interoperability, data integration, SPARQL, RDF

Presenting author

Marcos Zárate

Presented at

TDWG 2020

Funding program

This work is funded by the GBIF Young Researchers award 2019 granted to Marcos Zárate.

References

- Heath T, Bizer C (2011) Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology 1 (1): 1-136. https://doi.org/10.2200/s00334ed1v01y201102wbe001
- Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE 9 (8). https://doi.org/10.1371/journal.pone.0102623

- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715
- Zárate M, Braun G, Fillottrani P, Delrieux C, Lewis M (2020) BiGe-Onto: An ontology-based system for managing biodiversity and biogeography data1. Applied Ontology 1-27. https://doi.org/10.3233/ao-200228