Adding Biodiversity Datasets from Argentinian Patagonia to the Web of Data

Marcos Zárate[1,2,4] German Braun[3,4] Pablo Fillottrani[5,6]

[1] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico (CESIMAR-CENPAT)
[2] Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB)
[3] Universidad Nacional del Comahue (UNCOMA)
[4] Consejo Nacional de Invenstigaciones Científicas y Técnicas (CONICET)
[5] Computer Science and Engieneering Department, Universidad Nacional del Sur
[6] Comisión de Investigaciones Científicas de la provincia de Buenos Aires (CIC)

Abstract This poster presents a framework to publish biodiversity data from Argentinian Patagonia as Linked Open Data (LOD). These datasets contains information of biological species (mammals, plants, parasites, among others) that have been collected by researchers from the Centro Nacional Patagónico (CENPAT), and have initially been made available as Darwin Core Archive (DwC-A).

Keywords: Biocollections, Darwin Core, Linked Open Data, RDF

1 Introduction

Animal, plant and marine biodiversity comprise the “natural capital” that keeps our ecosystems functional and economies productive. However, since the world is experiencing a dramatic loss of biodiversity an analysis about this impact is being done by digitising and publishing biological collections. To this end, the biodiversity community has standardised shared common vocabularies such as *Darwin Core* (DwC) [1] together with platforms as the *Integrated Publishing Toolkit* (IPT) [2] aiming at publishing and sharing biodiversity data. As a consequence, the biodiversity community now have hundreds of millions of records published in common formats and aggregated into centralised portals. Therefore, new challenges emerge from this initiative for effectively using such a large volume of data. In particular, as the numbers of species, geographic regions, and institutions continue growing, answering questions about the complex interrelationships among these data becomes increasingly difficult. The Semantic Web (SW) [3] provides possible solutions to these problems by enabling the Web of Linked Data (LD) [4], where data objects are uniquely identified and the relationships amongst them are explicitly defined. LD is a powerful and compelling approach for spreading and consuming scientific data. It involves publishing, sharing, and connecting data on the Web, and offers a new way of data integration and interoperability. Moreover, there is an increasing recognition of the advantages of LD technologies in the life sciences. In this same direction, CENPAT[1] has started to publicly share its data under Open Data licence[2] through the IPT.[3] Data are available as Darwin Core Archive (DwC-A), which is a biodiversity data standard that makes use of the DwC terms, it is composed of a set of files for describing the structure and relationships of the raw data along with metadata files conforming the DwC standard. Nevertheless, the well-known IPT platform focuses on publishing content in unstructured or semi-structured formats but reducing the possibilities to interoperate with other datasets and make them accessible for machines. To enhance this approach, we present a transformation process from data extraction until its publishing as RDF datasets. This process uses OpenRefine[4] for generating RDF triples from semi-structured data and define URIs. It also uses GraphDB, for storing, browsing, accessing and link data with external RDF datasets.

2 Architecture Overview

Publishing data as LD involves data cleaning, mapping and a conversion process from DwC-A to RDF triples. The architecture of such a process is shown in Fig. 1 and has been structured as described below.

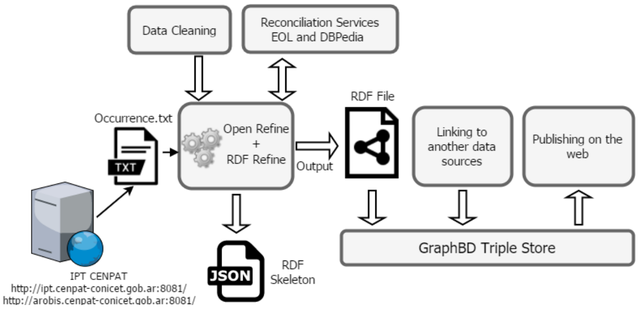
Data Cleaning
Reconciliation Services EOL and DBPedia
Occurrence.txt
TXT
Open Refine + RDF Refine
Output
RDF File
Linking to another data sources
Publishing on the web
GraphBD Triple Store
JSON
RDF Skeleton
IPT CENPAT
http://ipt.cenpat-conicet.gob.ar:8081/
http://arobis.cenpat-conicet.gob.ar:8081/

Figure 1. Transformation process for converting biodiversity datasets

(1) **Data Extraction, Cleaning and Reconciliation Process**, the DwC-A are manually extracted from the IPT repository and the files of occurrences are pre-processed (cleaning, conversion of data types, elimination of null values, etc.) using OpenRefine tool. OpenRefine also allows adding reconciliation services based on SPARQL endpoints, which return candidate resources belonging to external datasets for reconciling with fields from local datasets. In our process we use DBpedia endpoint to reconcile the `Country` column with the

[1] Patagonian National Research Centre, `http://www.cenpat-conicet.gob.ar/`

[2] `https://creativecommons.org/licenses/by/4.0/legalcode`

[3] `https://www.gbif.org/ipt`

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

`dbo:country` resource in DBpedia. Another reconciliation service was provided by *Encyclopedia of Life* (EOL)[5] which allows to reconcile taxonomic names. This service is applied to the `scientificName` column to obtain the URL of the EOL page describing the specie. (2) **RDF Schema Alignment and URI Definition**, after data cleaning and reconciliation, data are converted to RDF triples using RDF Refine.[6] The RDF schema alignment skeleton specifies the subject, predicate and the object of the triples to be generated. The next step is to set up prefixes for well-known vocabularies such as the W3C Basic Geo ontology, DBpedia, FOAF, DwC and Darwin-SW to establish relationships between DwC classes. Each resource must have an URI link that resource to other resources both within this dataset and others anywhere on the web. The common base URI for all the resources we define is `http://crowd.fi.uncoma.edu.ar:3333/`. (3) **Interlinking**, OpenRefine reconciliation service is able to match some links to DBpedia, but since it is still limited, our process should use a more powerful tool to discover links to other datasets. In this context, SILK[7] offers a graphical editor that can be used to create linkage rules. For example, the links to DBpedia have been generated taking into account the genus of the species, described by the term `dwc:genus` of the DwC and `dbo:genus` in DBpedia. The links between our RDF and DBpedia use the `owl:sameAs` predicate to link the two datasets. (4) **Publishing and Accessing Data**, the transformed biodiversity data have been published, and can to be accessed through GraphDB[8] allowing users to explore the hierarchy of RDF classes (`Class hierarchy`), similarly relationships among these classes also can be explored giving an overview about how many links exist between instances of the two classes (`Class relationship`).

3 Case Study: Conservation Status of Species

In this section we present a simple SPARQL query[9] to determine the conservation status of the species in our dataset. Since this information is not present, we can obtain it from the links to DBpedia using the property `owl:sameAs`, in this way our dataset benefits from information that did not previously exist.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX txn: <http://lod.taxonconcept.org/ontology/txn.owl#>

SELECT ?scname ?eol_page ?c_status
WHERE { ?s a dwc:Taxon.
        ?s dwc:scientificName ?scname.
        ?s txn:hasEOLPage ?eol_page.
        ?s owl:sameAs ?resource .
SERVICE <http://dbpedia.org/sparql> {
        ?resource dbo:conservationStatus ?c_status.}}
```

[5] http://www.eol.org/

[6] http://refine.deri.ie/

[7] http://silkframework.org/

[8] http://crowd.fi.uncoma.edu.ar:3333/

[9] http://crowd.fi.uncoma.edu.ar:3333/sparql?savedQueryName=c-status

4 Results

In order to test our architecture we use the datasets belonging to CENPAT, which are available as DwC-A in an institutional IPT server. These datasets include collections of marine, terrestrial, parasites and plant species mainly from Argentinean Patagonia. Up to July 2017, CENPAT owns 33 datasets representing about 273.419 occurrence records, where 80% of them have been georeferenced. In this initial stage only three datasets were converted to RDF, our platform stored 202.119 RDF triples. Also for the user to be able to exploit the dataset we define some SPARQL queries and their corresponding visualisation, for this we use the statistical software R, the scripts can be downloaded from[10] and a complete description of the proposed architecture can be found in.[11]

5 Conclusion and Outlook

In this poster we have presented the CENPAT Linked Open Biodiversity dataset, which exposes public biodiversity data related mainly to species from Argentinean Patagonia as LOD. The aim is to facilitate the access of researchers to important data and thus giving a valuable support to the scientific analysis of the biodiversity. In addition, this work is the first Argentinian initiative to convert biodiversity data according to the criteria established by LOD.

Finally, our approach have some limitations that we consider as future work. We need provide more advanced options and support automated execution of the extraction and conversion pipelines using, for example, *LinkedPipes ETL*[12] since this process is currently done manually.

References

1. 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.
2. 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*, 9(8):e102623, 2014.
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. *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227, 2009.

[10] `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/tree/master/r-scripts`

[11] `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/wiki`

[12] `https://etl.linkedpipes.com/`