Adding Biodiversity Datasets from Argentinian Patagonia to the Web of Data

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

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

Abstract In this work we present 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) have been collected by researchers from the Centro Nacional Patagónico (CENPAT), and have initially been made available as Darwin Core Archive (DwC-A) files. We introduce and detail a transformation process and explain how to access and exploit them, promoting integration with other repositories.

Keywords: Biocollections, Darwin Core, Linked data, RDF, SPARQL

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 [1,2], an analysis about its impact is being done by digitising and publishing biological collections [3]. To this end, the biodiversity community has standardised shared common vocabularies such as Darwin Core (DwC) [4] together with platforms as the Integrated Publishing Toolkit (IPT) [5] 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. Nevertheless, new challenges emerged from this initiative for effectively using such a large volume of data. In particular, as the number of species, geographic regions, and institutions continue growing, answering questions about the complex interrelationships among these data become increasingly difficult. The Semantic Web (SW) [6] provides possible solutions to these problems by enabling the Web of Linked Data (LD) [7], where data objects are uniquely identified and the relationships among 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. The driving force to implement LD spaces is the RDF technology. Moreover, there is an increasing recognition of the advantages of LD technologies in the life sciences [8,9].

In this same direction, CENPAT[1] has started to publicly share its data under Open Data licence.[2] Data are available as Darwin Core Archive (DwC-A) [10], which are 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 to publish these data as RDF datasets. This process uses OpenRefine [11] for generating RDF triples from semi-structured data and define URIs. It also uses GraphDB [12], previously known as OWLIM [12], for storing, browsing, accessing and linking data with external RDF datasets. Along this process, we follow the stages defined in the LOD Life-Cycle proposed in [13]. We claim that this work is an opportunity to exploit data from biodiversity in Argentina because they had been never published as LOD.

This work is structured as follows. Section 2 describes the main features of the datasets selected and their relationships with DwC. Section 3 describes the transformation process to RDF, while section 4 presents its publication and its access. Section 5 shows the framework to discover links to other datasets. Next, section 6 presents the exploitation of the dataset. Finally, we draw conclusions and suggest some future improvements.

2 CENPAT Data Sources

In this section, before describing our datasets, we briefly explain the DwC standard and DwC-A, which these datasets are based on.

2.1 Darwin Core Terms and Darwin Core Archive

DwC [4] is a body of standards for biodiversity informatics. It provides stable terms and vocabularies for sharing biodiversity data. DwC is maintained by TDWG[3] (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases). Its terms are organised into nine categories (often referred to as *classes*), six of which cover broad aspects of the biodiversity domain. *Occurrence* refers to existence of an organism at both particular place and time. *Location* is the place where the organism were observed (normally a geographical region or place). *Event* is the relationship between *Occurrence* and *Location* and register protocols and methods, dates, time and field notes.

---

[1] http://www.cenpat-conicet.gob.ar/
[2] https://creativecommons.org/licenses/by/4.0/legalcode
[3] http://www.tdwg.org/

Finally, *Taxon* refers to scientific names, vernacular names, etc. of the organism observed. The remaining categories cover relationships to other resources, measurements, and generic information about records. DwC also makes use of Dublin Core terms [14], for example: *type, modified, language, rights, rightsHolder, accessRights, bibliographicCitation, references.*

In the same direction, Darwin Core Archive (DwC-A) [10] is a biodiversity informatics data standard that makes use of the DwC terms to produce a single, self-contained dataset and thus sharing both species-level (taxonomic) and species-occurrence data. Moreover, each DwC-A includes these files. Firstly, the `core data file` (mandatory) consists of a standard set of DwC terms together with the raw data. This file is formatted as fielded text, where data records are expressed as rows of text, and data elements (columns) are separated with a standard delimiter such as a tab or comma. Its first row specifies the headers for each column. Secondly, the `descriptor metafile` defines how the core data file is organised and maps each data column to a corresponding DwC term. Lastly, the `resource metadata` provides information about the dataset itself such as its description (abstract), agents responsible for authorship, publication and documentation, bibliographic and citation information, collection method, among others.

2.2 Dataset Features

The datasets analysed belong to CENPAT and are available as DwC-A in an IPT server from this institution. They include collections of marine, terrestrial, parasites and plant species mainly registered from several points of the Argentinian Patagonia. Data are generated in different ways: some of them by means of electronic devices placed in different animals to study environmental variables, while others are observations of species in their natural habitat or species studied in laboratories. To ensure the quality of these data, the records have been structured according to the procedure described in [15].

Up to May 2017, CENPAT owns 33 datasets representing about 273.419 occurrence records, where 80% of them have been also georeferenced. Some of these collections contain unique data never published because of the age of the records (1970s). As a consequence, making this information available as LOD is so important for researchers, who are studying species conservation and the impact of man in biodiversity along the last years [16,17].

3 Linked Data Creation

Publishing data as LD involves data cleaning, mapping and conversion processes from DwC-A to RDF triples. The architecture of such a process is shown in Fig. 1 and has been structured as described in the following subsections.
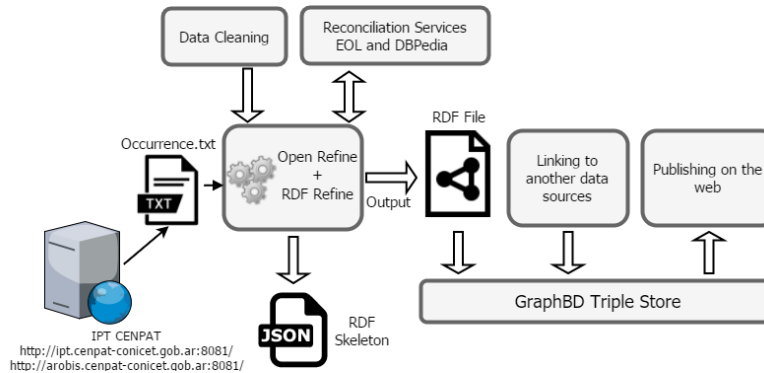
Figure 1. Transformation process for converting biodiversity datasets

3.1 Data Extraction, Cleaning and Reconciliation Process

The DwC-A are manually extracted from the IPT repository and their occurrences files (`occurrence.txt`) are processed using OpenRefine tool [11]. There, occurrences are cleaned and converted to standardised data types such as dates, numerical values, etc. and empty columns are removed. OpenRefine also allows adding reconciliation services based on SPARQL endpoints, which return candidate resources from external datasets to be matched to fields in the local datasets. In our process, we use DBpedia [18] endpoint[4] to reconcile the `Country` column with the `dbo:country` resource in DBpedia, the link between the resources is made through the property `owl:sameAs`. After that, if the reconciliation is done, we create a new column for the corresponding URI of the resource. In particular, we add the column named `dbpediaCountryURI` for the original `Country`.

Another reconciliation service[5] used, it was based on a taxonomic database *Encyclopedia of Life* (EOL)[6] which allows to reconcile *accepted names* in EOL database. Specifically, the reconciliation is applied to the column `scientificName` so that we create a new column named `EOL_page` for the EOL page describing the specie. Unfortunately, this whole process is time-consuming because not all values are automatically matched and thus ambiguous suggestions must be fixed. Moreover, in this phase only two columns have been possible to reconcile because the process returns unsuitable results using DBpedia services some columns like `institutionCode` or `locality`.

---

[4] https://dbpedia.org/sparql
[5] http://iphylo.org/~rpage/phyloinformatics/services/reconciliation_eol.php
[6] http://www.eol.org/

3.2 RDF Schema Alignment and URI Definition

After cleaning and reconciling, data are converted to RDF triples using RDF Refine[7], which is an extension of OpenRefine tool. RDF Refine allows users to go through a graphical interface describing the RDF scheme alignment skeleton to be shared among different datasets. The RDF skeleton 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 the W3C Basic Geo ontology [19], Geonames [20], DBpedia, FOAF [21], Darwin-SW [22] for establishing relationships among DwC classes and Taxon Concept.[8] Table 1 shows the prefixes used.

Table 1. Prefix used in the mapping process.

Prefix	Description	URI
cnp-gilia	Base URI	http://crowd.fi.uncoma.edu.ar:3333/
dwc	Darwin Core	http://rs.tdwg.org/dwc/terms/
dws	Darwin-SW	http://purl.org/dsw/
foaf	Friend of a Friend	http://xmlns.com/foaf/0.1/
dc	Dublic Core	http://purl.org/dc/terms/
geo-pos	WGS84 lat/long vocab	http://www.w3.org/2003/01/geo/wgs84_pos#
geo-ont	GeoNames	http://www.geonames.org/ontology#
wd	Entitys in Wikidata	http://www.wikidata.org/entity/
wdt	Properties in Wikidata	http://www.wikidata.org/prop/direct/
txn	Taxon Concept Ontology	http://lod.taxonconcept.org/ontology/txn.owl#

In order to generate URI for each resource, in this approach we used GREL (General Refine Expression Language) also provided by OpenRefine, the general structure of the URIs is described below:

$$\texttt{http://[base\_uri]/[DwC class]/[value]}$$

where: [base_uri] is the one specifies in Table 1, [DwC class] is the respective DwC class and [value] is the value of the cells in the file of occurrences. It is also important to note that the generated URIs are instances of the classes defined in the DwC standard. Finally, the resulting RDF triple for an occurrence is:

```
SUBJECT: <base_uri/occurrence/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
PREDICATE: rdf:type
OBJECT: dwc:Occurrence
```

Table 2 describes the mapping performed and which columns have been used to generate the main URIs.

---

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

[8] http://lod.taxonconcept.org/ontology/txn.owl

Table 2. The first part of the table shows the main classes corresponding to the categories of the DwC standard. Moreover, the columns of the DwC-A file used to generate URIs. The second part shows the properties used and an example of the literals obtained from the columns of the file of `occurrences.txt`. For simplicity, the table shows only the main properties, see the complete scheme at `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/Open_refine_scripts/rdf_skelton.json`

Class	Columns used to create URI	URI example
dwc:Taxon	genus + specificEpithet	<base_uri:taxon/Mirounga_leonina>
dwc:Occurrence	id	<base_uri:occurrence/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
dwc:Event	id	<base_uri:event/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
dwc:Dataset	dataset	<base_uri:dataset/dwca-mamcenpat-v1.1>
dc:Location	id	<base_uri:location/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
foaf:Agent	institutionCode	<base_uri:agent/cenpat-conicet>

Property	Columns used	Example
dwc:class	class	"Mammalia"^^xsd:string
dwc:family	family	"Phocidae"^^xsd:string
dwc:genus	genus	"Mirounga"^^xsd:string
dwc:kingdom	kingdom	"Animalia"^^xsd:string
dwc:order	order	"Carnivora"^^xsd:string
dwc:phylum	phylum	"Chordata"^^xsd:string
dwc:scientificName	scientificName	"Mirounga leonina Linnaeus, 1758"^^xsd:string
txn:hasEOLPage	EOL_page	"http://eol.org/pages/328639"^^xsd:string
dwc:basisOfRecord	basisOfRecord	"PreservedSpecimen"^^xsd:string
dwc:occurrenceRemarks	occurrenceRemarks	"craneo completo"^^xsd:string
dwc:individualCount	individualCount	1^^xsd:int
dwc:CatalogNumber	CatalogNumber	"100751-1"^^xsd:string
geo-pos:lat	decimalLatitude	-42.53^^xsd:decimal
geo-pos:long	decimalLongitude	-63.6^^xsd:decimal
geo-ont:countryCode	country	"Argentina"^^xsd:string
dwc:verbatimEventDate	dwc:verbatimEventDate	"2004-10-22"^^xsd:date
foaf:name	recordedBy or InstitutionCode	"CENPAT-CONICET"@en.

4 Publishing and Accessing Data

The transformed biodiversity data have been published, and can to be accessed, through GraphDB. GraphDB is a highly efficient and robust graph database with RDF and SPARQL support. It allows users to explore the hierarchy of RDF classes (`Class hierarchy`), where each class can be browsed to explore its instances. 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`). Each link is a RDF statement where its subject and object are class instances and its predicate is the link itself. Lastly, users also can explore resources providing URIs representing any of the subject, predicate or object of a triple (`View resource`).

Finally, Fig. 2 shows the resulting graph for the description of a *southern elephant seal skull*, which is part of the CENPAT collection of marine mammals and contains information about where has been found, who has been collected for, sex and scientific name, among others. Another way to access the same information is to explore the `View resource` in the GraphDB repository `http://crowd.fi.uncoma.edu.ar:3333/resource/find` for the specific occurrence `f6bbf85d-85ea-4605-87fa-d81aca73a1cd`, while the serialization of the complete graph in Turtle syntax can be consulted in.[9]
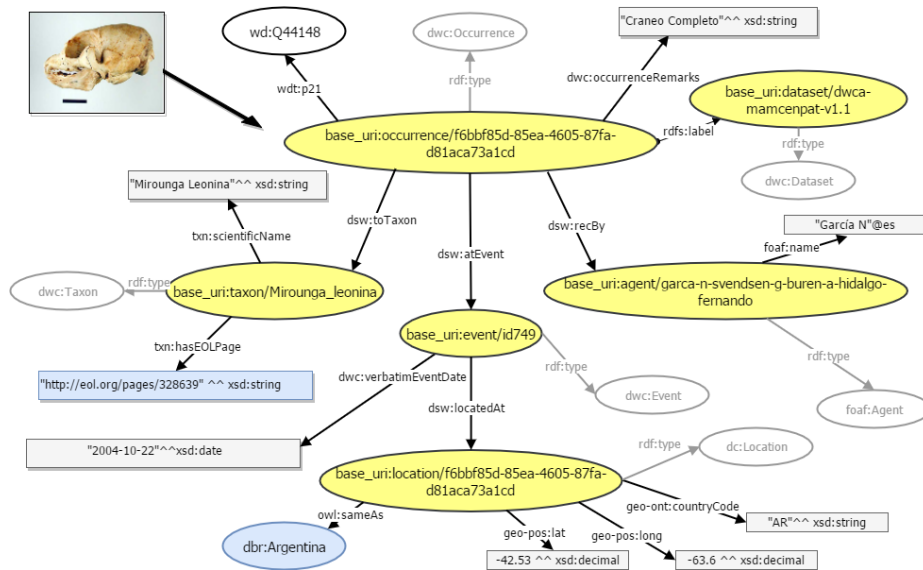

Figure 2. Figure shows links between instances of classes, `rdf:type` assertions are shown in light gray. In blue colour you can see the reconciled values .

---

[9] `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/rdf/graph.ttl`, accessed at September 2017

5 Interlinking

Interlinking other datasets in a semi-automated way is crucial aiming at facilitating data integration. In this context, OpenRefine reconciliation service is able to match some links to DBpedia, but since it is still limited, our process should use more powerful tools to discover links to other datasets. For this task, our approach preliminarily integrate SILK framework[10] that uses `Silk-Link Specification Language` (Silk-LSL) to express heuristics for deciding whether a semantic relationship exists between two entities. For interlinking species between DBpedia and our dataset, we used *Levenshtein distance* a comparison operator that evaluates two inputs and computes the similarity based on a user-defined distance measure and a user-defined threshold. This comparator receives as input two strings `dbp:binomial` (Binomial nomenclature in DBpedia) and the combination of `dwc:genus` + `dwc:specificEpithet` (the concatenation of these two defines the scientific name of the species). The Levenshtein distance comparator was set up with `<Thresholds = "0.0" and Weight = "1">`. After the execution, SILK discovered 15 links to DBpedia with an accuracy of 100% and 85 link with an accuracy between 65% and 75%. In this case, we permit only one outgoing `owl:sameAs` link from each resource. The complete Silk-LSL script can be downloaded from.[11]

However, although a set of links has been successfully generated, users' feedback is needed to filter some species wrongly matched by the tool. Finally, we must identify further candidates for interlinking and tests other properties or classes from our dataset in order to increase the automatic capabilities of the framework.

6 Exploitation

This section shows how the different types of observations of species can be retrieved, complemented with information of another datasets and filtered by submitting SPARQL queries to GraphDB endpoint. Moreover, it provides some experiments in R by using the `SPARQL`[12] package. Each SPARQL query in following examples assumes the prefix defined in Table 1.

Total Number of Species in the CENPAT Dataset. The following query retrieves the species of the dataset. To this end, it includes the scientific name of the species and also its amount of occurrences, to execute this query in GraphDB see.[13] The Fig. 3 shows only the first resulting records.

---

[10] `http://silkframework.org/`

[11] `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/SILK/link-spec.xml`, accessed at September 2017

[12] `https://cran.rproject.org/web/packages/SPARQL/SPARQL.pdf`

[13] `http://crowd.fi.uncoma.edu.ar:3333/sparql?savedQueryName=species-count`

```
SELECT ?scname (COUNT(?s) AS ?observations)
      {?s a dwc:Occurrence.
       ?s dsw:toTaxon ?taxon.
       ?taxon dwc:scientificName ?scname }
GROUP BY ?scname
ORDER BY DESC(COUNT(?s))
```

```
--------------------------------------------------------
scname                                        observations
--------------------------------------------------------
Otaria flavescens                                    2,787
Mirounga leonina                                     1,851
Merluccius hubbsi                                      368
```

Figure 3. Occurrences of each species that contains the dataset.

Occurrences by Year. The following query allows to observe the temporality of the occurrences and its results are visualised using R as shown the Fig. 4. The R script is available in.[14]

```
SELECT ?year (COUNT(?s) as ?count)
      {?s a dwc:Event.
       ?s dwc:verbatimEventDate ?date }
GROUP BY (year(?date) AS ?year)
ORDER BY ASC(?year)
```
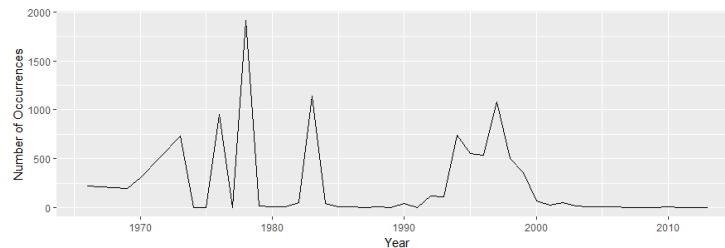

Figure 4. Simple plot using SPARQL and ggplot2 package for R.

Conservation Status of Species. Conservation status are defined by The IUCN Global Species Programme[15] and are taken as a global reference. Information about the state of conservation is missing in CENPAT datasets so that

---

[14] https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/
r-scripts/occurrences-by-year.R, accessed at September 2017
[15] http://www.iucnredlist.org/

providing these data linking other RDF datasets is highly desirable. To this end, the following query capture these missing data using the `owl:sameAs` property. The results are shown in Fig. 5, to execute this query in GraphDB, see.[16]

```
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.}
}
```

```
===========================================================================
scname                                            eol_page               c_status
---------------------------------------------------------------------------
Otaria flavescens                                 http://eol.org/pages/328614    LC
Mirounga leonina                                  http://eol.org/pages/328639    LC
Hydrurga leptonyx Blainville, 1820                http://eol.org/pages/328637    LC
Arctocephalus australis                           http://eol.org/pages/328623    LC
Lagenorhynchus obscurus Gray, 1828                http://eol.org/pages/317317    DD
Phocoena dioptrica Lahille, 1912                  http://eol.org/pages/328461    DD
Hyperoodon planifrons Flower, 1882                http://eol.org/pages/328557    LC
Mesoplodon grayi Von Haast, 1876                  http://eol.org/pages/328562    DD
```

Figure 5. Conservation status associated to the species: LC (Least Concern), DD (Data Deficient), EN (Endangered), VU (Vulnerable).

Locations of Marine Mammals. The last query is to retrieve the locations (latitude and longitude) for the species *Mirounga Leonina*. The results are depicted in Fig. 6 using R, and the script is available in.[17]

```
SELECT ?lat ?long
WHERE { ?s a dwc:Occurrence.
        ?s dsw:toTaxon ?taxon.
        ?taxon dwc:scientificName ?s_name.
        ?s dsw:atEvent ?event.
        ?event dsw:locatedAt ?loc.
        ?loc geo-pos:lat  ?lat .
        ?loc geo-pos:long ?long
FILTER (?lat >= "-58.4046"^^xsd:decimal && ?lat <= "-32.4483"^^xsd:decimal)
FILTER (?long >= "-69.6095"^^xsd:decimal && ?long <= "-52.631"^^xsd:decimal)
FILTER regex ( STR (?s_name ), "Mirounga␣leonina")}
```

7 Conclusions and Further Works

In this work we have presented a framework to publish biodiversity data from Argentinian Patagonia as LOD, which have initially been made available as

---

[16] http://crowd.fi.uncoma.edu.ar:3333/sparql?savedQueryName=
conservation-status
[17] https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/
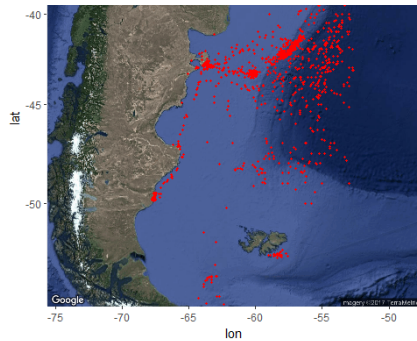r-scripts/positions-ml.R, accessed at September 2017

Figure 6. Visualization of animal movements using R

Darwin Core Archive files. 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.

We have detailed the transformation process and explained how to access and exploit them, promoting integration with other repositories. Moreover, we have depicted this process using queries extracted from the domain of application. Such RDF repository is hosted at `http://crowd.fi.uncoma.edu.ar:3333/` together with an SPARQL endpoint, in this initial stage we store 202.119 triples.

As future works, we plan to automate some tasks of the process and interlink with more datasets. Moreover, providing easier SPARQL access for non-skilled users. Finally, we are analysing other ontologies such as ENVO [23], NCBI [24] and OWL Time [25] and working on a suite of complementary ontologies for describing every aspect of semantic biodiversity.

References

1. Craig Moritz, James L Patton, Chris J Conroy, Juan L Parra, Gary C White, and Steven R Beissinger. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. *Science*, 2008.
2. Adriana Vergés, Peter D Steinberg, Mark E Hay, Alistair GB Poore, Alexandra H Campbell, Enric Ballesteros, Kenneth L Heck, David J Booth, Melinda A Coleman, and Feary. The tropicalization of temperate marine ecosystems: climate-mediated changes in herbivory and community phase shifts. In *Proc. R. Soc. B*. The Royal Society, 2014.
3. Malcolm Scoble. Rationale and value of natural history collections digitisation. *Biodiversity Informatics*, 2010.
4. 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.
5. 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*, 2014.

6. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The Semantic Web. *Scientific American*, 2001.

7. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, 2009.

8. François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 2008.

9. Jouni Tuominen, Nina Laurenne, and Eero Hyvönen. *Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception*. Springer, 2011.

10. K Döring M Robertson T Remsen D, Braak. Darwin Core Archive How-To Guide. 2011.

11. Ruben Verborgh and Max De Wilde. *Using OpenRefine*. Packt Publishing Ltd, 2013.

12. Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, and Ruslan Velkov. OWLIM: A family of scalable semantic repositories. *Semantic Web*, 2011.

13. Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N. Mendes, Bert Van Nuffelen, Claus Stadler, Sebastian Tramp, and Hugh Williams. Managing the Life-Cycle of Linked Data with the LOD2 Stack. In *International Semantic Web Conference (2)*, Lecture Notes in Computer Science, 2012.

14. Dublin Core Metadata Initiative et al. Dublin core metadata element set, version 1.1. 2012.

15. Mark J Costello and John Wieczorek. Best practice for biodiversity data management and publication. *Biological Conservation*, 2014.

16. Reed S Beaman and Nico Cellinese. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, 2012.

17. Ana Vollmar, James Alexander Macklin, and Linda Ford. Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics*, 2010.

18. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web*, 2007.

19. D Brickley. W3C Semantic Web Interest Group: Basic Geo (WGS84 lat/long) Vocabulary, 2011.

20. Marc Wick, B Vatant, and B Christophe. Geonames ontology. *http://www.geonames.org/ontology, accessed at Sep 2017*, 2015.

21. Dan Brickley and Libby Miller. The Friend Of A Friend (FOAF) vocabulary specification, 2007.

22. Steven J Baskauf and Campbell O Webb. Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. *Semantic Web*, 2016.

23. Pier Luigi Buttigieg, Evangelos Pafilis, Suzanna E. Lewis, Mark P. Schildhauer, Ramona L. Walls, and Christopher J. Mungall. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, 2016.

24. Scott Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 2012.

25. Time Ontology in OWL, 2006. `http://www.w3.org/TR/owl-time`, accessed at September 2017.