UNIVERSIDAD NACIONAL DEL SUR

Tesis de Doctorado en Ciencias de la Computación

Alineamiento e integración de información basada en ontologías para Biogeografía marina y Biodiversidad

Marcos Daniel Zárate

Bahía Blanca Argentina

2019

Prefacio

Esta Tesis se presenta como parte de los requisitos para optar al grado académico de Doctor en Ciencias de la Computación, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Laboratorio de Investigación & Desarrollo en Ingeniería de Software y Sistemas de Información (LISSI), durante el período comprendido entre el 3 de Noviembre de 2015 y el 24 de Octubre de 2019, bajo la dirección del Dr. Pablo Fillottrani, Profesor Titular del Departamento de Ciencias e Ingeniería de la Computación.

Marcos Daniel Zárate
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Bahía Blanca, 24 de Octubre de 2019.

Agradecimientos

Quiero comenzar por agradecer a mi director Dr. Pablo Fillottrani, por la paciencia que tuvo para guiarme, a él debo agradecer el empuje, la motivación que me dio para llevar adelante esta investigación. A la Dra. Mirtha Lewis, del CESIMAR-CENPAT-CONICET, quiero agradecer el ánimo brindado y sus aportes permanentes a mis trabajos. Al Dr. Claudio Delrieux, del DIEC-UNS, quiero agradecer su predisposición permanente para responder mis consultas y todas las charlas que mantuvimos, las cuales me hicieron crecer como tesista.

Estoy en deuda con todos mis compañeros docentes de la Facultad de ingeniería de la Universidad Nacional de la Patagonia San Juan Bosco, por ser parte de mi formación de grado y por fomentar mi interés por las ciencias de la computación. Quiero agradecer en particular al Lic. Renato Mazzanti y al Lic. Carlos Buckle por apoyarme incondicionalmente. Gracias a mis compañeros del CESIMAR-CENPAT-CONICET, por el alegre entorno de trabajo y por las discusiones enriquecedoras que tenemos en forma permanente.

A mis padres deseo agradecerles por hacer lo imposible para que sea una persona de bien y por ser dos referentes que admiro y amo profundamente. Finalmente gracias a todos mis amigos y a mi compañera Dana, por ser un soporte fundamental en mi vida.

Resumen

El objetivo principal de esta tesis es analizar los problemas que existen actualmente con el manejo integrado de información en las ciencias de la vida en general, y particularmente analizar que sucede con la Biodiversidad y la Oceanografía. La actual crisis mundial de la biodiversidad, debida, entre otras cosas, al calentamiento global, genera un profundo impacto en la distribución geográfica de las especies y las comunidades ecológicas. Esto provoca un creciente interés entre los científicos para coordinar el uso compartido de conjuntos de datos que ayuden a entender esta problemática global. En este contexto, el paradigma de los Datos Vinculados (Linked Data en inglés) ha emergido como un conjunto de buenas prácticas para conectar, compartir y exponer datos y conocimiento, una parte central de este paradigma son las ontologías, que permiten la definición de vocabularios compartidos y modelos conceptuales que ayuden a integrar esta información. Estas consideraciones proporcionan una fuerte motivación para formular un sistema que tenga en cuenta las características geoespaciales que pueden brindar respuestas a preguntas como las siguientes: (i) *¿Cómo podemos definir las regiones espaciales para nuestros estudios?* (ii) *¿Cómo se distribuyen las especies en una determinada región?* (iii) *Dada una georeferencia particular, ¿a qué región geográfica pertenece?* (iv) *¿Cómo relacionar las ocurrencias de especies con variables ambientales dentro de una región específica?*.

En esta tesis se presenta el desarrollo de un sistema basado en ontologías denominado *BiGe-Onto* [ZBF+19] para administrar información de los dominios de Biodiversidad y Biogeografía marina. Este sistema está compuesto por (i) Arquitectura; (ii) Modelo conceptual; (iii) Versión operacional OWL 2; y (iv) Conjunto de datos vinculados para su explotación a través de un punto final SPARQL.

La evaluación de *BiGe-Onto* se realizo desde dos enfoques, el primero de ellos consiste en validar la ontología utilizando datos reales extraídos de repositorios de Biodiversidad y Biogeografía marina para luego validar el modelo conceptual propuesto utilizando preguntas de competencia. El segundo enfoque tiene que ver con la validación mediante casos de estudio

definidos en conjunto con investigadores del Centro Científico Tecnológico (CENPAT-CONICET) que trabajan realizando análisis de distribución de especies. Finalmente la documentación de *BiGe-Onto* esta disponible en línea en http://crowd.fi.uncoma.edu.ar/cenpat-gilia/bigeonto/ y el conjunto de datos enlazados es accesible públicamente a través de DOI 10.5281/zenodo.3235548.

Abstract

The main goal of this thesis is to analyze the existing issues currently related to the integrated management of information in life sciences in general, and particularly to analyze what happens with Biodiversity and Oceanography. The current global biodiversity crisis, due, among other things, to global warming, has a great impact on the geographical distribution of species and ecological communities. This motivates a growing interest among scientists to coordinate the sharing of datasets that help to understand this global problem. In this context, Linked Data paradigm has emerged as a set of good practices to connect, share and expose data and knowledge. A central part of this paradigm are the ontologies, which allow the definition of shared vocabularies and conceptual models that help integrate this information. These considerations provide strong motivation to formulate an ontology-based system considering geospatial features that may provide answers to questions such as: **(i)** How can we define spatial regions for our studies? **(ii)** How are the species distributed in a certain region? **(iii)** Given a particular georeference, which geographic region does it belong to? **(iv)** How to relate occurrences of species with environmental variables within a specific region?.

This thesis presents the development of an ontology-based system called *BiGe-Onto* [ZBF+19] to manage information from Biodiversity and Marine Biogeography domains. This system is composed of (i) Architecture; (ii) Conceptual model; (iii) OWL 2 operational version; and (iv) Linked dataset to exploit through a SPARQL endpoint. *BiGe-Onto* evaluation was developed from two approaches, the first one is to validate the ontology using real data extracted from Biodiversity and Marine Biogeography repositories and then validate the proposed conceptual model using competence questions. The second approach is based on validation through case studies defined in conjunction with researchers from the Technological Scientific Center (CENPAT-CONICET) who work on species distribution analysis. Finally, *BiGe-Onto* documentation is available online at `http://crowd.fi.uncoma.edu.ar/cenpat-gilia/bigeonto/` and the linked dataset is publicly accessible through DOI 10.5281/zenodo.3235548.

Índice general

Capítulo 1

Introducción

Índice

En este capítulo, se introducen el contexto y la motivación que guiaron el desarrollo de esta investigación y que dieron como resultado el desarrollo de un modelo conceptual y la publicación de un conjunto de datos abiertos enlazados que permiten el manejo integrado de información de Biogeografía marina y Biodiversidad teniendo en cuenta las características geoespaciales. Asimismo, se sintetizan las principales contribuciones que se han obtenido y que han sido publicadas en revistas y o congresos nacionales e internacionales. Finalmente, se resume la organización de esta Tesis.

1.1. Motivación

La cantidad de información que se genera en la Web se ha incrementado en los últimos años. La mayor parte de esta información se encuentra accesible en texto, ya que el principal usuario de la Web es el ser humano. Sin embargo, a pesar de todos los avances producidos en el área del procesamiento del lenguaje natural, los ordenadores tienen problemas para procesar esta información textual [RJS11, GRP+16]. Sin bien, existen dominios de aplicación en los que se están publicando grandes cantidades de información como datos estructurados[1]. En el caso general de las Ciencias de la Vida y en particular la Oceanografía y la Biodiversidad, han generado una enorme cantidad de datos estructurados durante la última década [Sin13]. El análisis de estos datos es de vital importancia no sólo para el avance de la ciencia, sino para producir avances en el estudio de los océanos y la conservación de las especies. Sin embargo, estos datos están localizados en diferentes repositorios y almacenados en diferentes formatos que hacen difícil su integración. En este contexto, el paradigma de los Datos Enlazados [HDHC14] ha emergido como un conjunto de buenas prácticas para conectar, compartir y exponer datos y conocimiento. Esta tecnología incluye la aplicación de algunos estándares propuestos por el World Wide Web Consortium (W3C) tales como HTTP URIs [W3c01], los estándares RDF [W3C04] y OWL [OWL09] y el lenguaje de consulta SPARQL [SPA08].

La comunidad de los Datos Abiertos Enlazados (Linked Open Data o LOD por sus siglas en inglés) ha fomentado la publicación de conjuntos de datos enlazados. El número de conjuntos de datos y tripletas RDF publicados en LOD se ha incrementado en esta última década. En 2007, los 12 conjuntos de datos que formaban la nube de datos enlazados contenían más de 2 billones de tripletas y 2 millones de documentos RDF. En el año 2013, la nube contenía 2.289 repositorios y más de 11 billones de tripletas atendiendo a las estadísticas oficiales publicadas. Actualmente (Mayo de 2019) el

[1] En este contexto cuando hablamos de datos estructurados nos referimos a archivos de tipo texto que se suelen mostrar en filas y columnas con títulos.

número de repositorios corresponde a 2.973 y el número de tripletas equivale a 140 billones. Las Ciencias de la Vida fue uno de los primeros campos en adoptar la tecnología de los Datos Enlazados. Esta adopción se ha materializado en una gran cantidad de datos en este campo que corresponde a un 11,05 % del total de la nube de Datos Enlazados. No obstante, a pesar de los esfuerzos en publicar datos usando estas tecnologías, existe cierta escasez de aplicaciones que recuperen información, proporcionen nuevas perspectivas a los Datos Enlazados ya publicados o generen nuevas asociaciones en RDF que pueden resultar útiles. Dado este contexto, se han identificado un conjunto de limitaciones en la tecnología de los Datos Enlazados en el dominio de las Ciencias de la Vida:

- La disponibilidad del conjunto de datos. Cuando se habla de disponibilidad, se hace referencia a que existen repositorios RDF que han dejado de funcionar debido a la falta de mantenimiento por parte de los proveedores de datos. Actualmente, nos encontramos con una clara diferenciación en los repositorios que no suelen aplicar técnicas que aseguren la disponibilidad de los mismos.
- La heterogeneidad semántica. Este problema se refiere al hecho de que cada repositorio RDF presenta diferentes vocabularios, URIs, modelos de datos, etc. Esto complica la integración de estos datos y su reutilización.
- Una curva de aprendizaje empinada, dado que los usuarios presentan dificultades para aprender esta tecnología.
- Publicación de modelos de datos (Ontologías, vocabularios utilizados en el modelado de los datos).

El primer problema derivado del uso de esta tecnología se refiere a la disponibilidad de los repositorios RDF. En la categoría de las Ciencias de la Vida, existen varios ejemplos. Bio2RDF [BNT+08] es uno de ellos, Bio2RDF es un proyecto que se constituyó en el año 2008 para integrar múltiples fuentes

de datos RDF (incluida el propio conjunto de datos Bio2RDF) que pueden ser consultados de manera integrada. No obstante, muchas de estas fuentes de datos han dejado de funcionar por lo que la idea original se convirtió en una expresión de deseo. Para solventar el problema, los desarrolladores de Bio2RDF han proporcionado una página basada en JavaScript que contiene volcados de datos que los usuarios pueden descargar. Sin embargo, la página no es fácil de analizar y los datos no se actualizan frecuentemente. Otro ejemplo de este problema en el contexto de las ciencias del mar es el repositorio de campañas oceanográficas llamado Rolling Deck to Repository (R2R) [ACS+13] que proporciona un endpoint siempre disponible pero su información no se ha actualizado en años.

El segundo problema se refiere a la heterogeneidad semántica existente en los repositorios RDF en el campo de Ciencias de la Vida, por ejemplo existen conceptos iguales expresados con URIs diferentes, es aquí donde se hace indispensable la necesidad de un proceso de reconciliación de URIs que no es posible solucionar mediante el uso de la federación de consultas SPARQL.

El tercer problema hace referencia a la acusada curva de aprendizaje por parte de los usuarios finales para entender la tecnología los Datos Enlazados. Esto provoca que el uso de los Datos Enlazados se limite a los desarrolladores de aplicaciones, reduciendo el impacto de esta tecnología. Es necesario, por tanto, producir soluciones que acerquen esta tecnología a los usuarios finales, ampliando su impacto a mediano/largo plazo.

Finalmente, el cuarto problema que se detectó en la aplicación de la tecnología de los Datos Enlazados fue la ausencia de los modelos semánticos en repositorios RDF. La construcción de repositorios RDF está guiada por el uso de un modelo semántico que proporciona todos los elementos que representan un grafo RDF (sección 2.2.1). Este modelo proporciona todas las clases y relaciones que son usadas para la descripción de los datos. Las consultas SPARQL usarán estos elementos para la extracción de datos. El desconocimiento del modelo de datos subyacente dificulta el diseño de consultas, que

en el caso de los Datos Abiertos Enlazados no realiza el propietario de los datos. La mayoría de las técnicas presentadas en la literatura se refieren a vocabularios y patrones que son aplicados a un conjunto reducido de repositorios. Una parte fundamental de los Datos Enlazados son las ontologías, las cuales han tenido numerosas aplicaciones en el campo de la Ciencias de la Vida. El estándar del W3C para ontologías (OWL) (sección 2.2.3), se basa en lógicas descriptivas. Esto proporciona a las ontologías y los datos (instancias) la capacidad de aplicar técnicas de razonamiento. Esta característica es aprovechada por la familia de aplicaciones software conocidas como razonadores. Sin embargo, y pese al amplio uso de las ontologías en las Ciencias de la Vida, esta capacidad de razonamiento normalmente no se aprovecha. Visto el razonamiento como un proceso que incluye la clasificación, su aplicación en Biodiversidad es muy interesante para clasificar especies, áreas geográficas, etc.

1.2. Objetivos y Metodología

Los datos enlazados desempeñan un papel clave para interconectar datos de diferentes fuentes mediante el uso de estándares para facilitar el procedimiento para publicar, compartir y reutilizar estos datos. En particular, lo datos enlazados facilitan la realización de muchos de los principios demoninados FAIR (por sus siglas en inglés de Findable Accessible Reusable Interoperable) [WDA+16] al proporcionar un enfoque inherente al procesamiento de las máquinas, identificadores para cada elemento y el soporte para combinar elementos de datos y su semántica. De acuerdo con el conjunto de problemas descritos anteriormente en la motivación, plantemos los siguientes objetivos:

- Desarrollar una ontología que permita el manejo integrado de información de Biodiversidad y Biogeografía marina.
- Se validará esta ontología utilizando un razonador OWL, previamente

poblada con datos (instancias) extraídos de Ocean Biogeographic Information System (OBIS) [OBI19] y Global Biodiversity Information Facility (GBIF) [GBI19].

- Utilizar la ontología para realizar controles semiautomáticos en la calidad de los datos georeferenciados.
- Validar la ontología con la comunidad de usuarios involucrados.
- Crear un entorno colaborativo que facilite el consumo de Datos Enlazados por usuarios finales a través de consultas federadas que recuperen información de más de un repositorio.
- Promover el uso de los Datos Enlazados en el ámbito de la Biodiversidad y Biogeografía marina.

Para llevar a cabo los objetivos mencionados, se completaron las siguientes fases:

- Análisis del estado del arte actual sobre estudios que aplican la tecnología de los Datos Enlazados a los dominios de estudio, se investigó el uso de ontologías y sistemas para administrar información integrada de ambos dominios.
- Diseño e implementación de la versión operacional de la ontología en OWL 2. Esta fase incluye la validación con un razonador. Finalmente, se incluye la validación con casos de uso reales para mostrar su aplicabilidad en el manejo integrado de los datos.
- Diseño e implementación de una técnica semiautomática para detectar errores en los datos georeferenciados.
- Todo el proceso fue documentado mediante el uso de bitácoras y reportes internos de avance.
- Publicación de resultados de las distintas soluciones en congresos (nacionales e internacionales) y revistas científicas afines.

Inicialmente el plan de investigación propuesto contemplaba integrar el dominio oceanográfico con el de la biodiversidad, luego de realizar un análisis detallado de este dominio y evaluar los conjuntos de datos a utilizar, concluimos que para el propósito de esta tesis era suficiente utilizar datos de Biogeografía marina debido a la disponibilidad inmediata de los datos y dado que también tiene en cuenta el análisis de variables, mediciones y procedimientos que son comunes al de la oceanografía.

1.3. Resultados esperados

Como se mencionó anteriormente existen diversos problemas en la ciencias de la vida en lo que respecta al manejo integrado de la información, de este estudio, análisis y motivación, confluye la contribución central de nuestra investigación: *El manejo integrado de información de Biodiversidad y Biogeografía marina*. En este sentido, podemos distinguir los siguientes aportes:

- Desarrollo y publicación un sistema basado en ontologías denominado *BiGe-Onto* que permite el manejo integrado de la información extraída de OBIS y GBIF. *BiGe-Onto* fue validada por expertos del dominio pertenecientes al Centro Nacional Patagónico (CCT CONICET–CENPAT)[2]. *BiGe-Onto* contribuye a acercar las tecnologías de los Datos Enlazados a expertos del dominio.
- Publicar un conjunto de Datos Abiertos Enlazados accesible a través de un punto final SPARQL, con información de Biodiversidad/Biogeografía, siguiendo las mejores practicas para la publicación de datos abiertos enlazados [BCH+07].
- Utilizar *BiGe-Onto* y el conjunto de datos integrado para detectar errores en las georeferencias provenientes de OBIS y GBIF. Para el manejo de los datos georeferenciados se integró la ontología GeoSPARQL [PH12] la cual permite realizar operaciones geoespaciales y

[2] http://www.cenpat-conicet.gob.ar/

razonar sobre los datos. Estos resultados muestran que *BiGe-Onto* no solo es una herramienta prometedora el manejo integrado de información, sino que sirve como herramienta pera el control de calidad de datos.

1.4. Organización de la Tesis

La tesis incluye los contenidos necesarios para que su lectura sea autocontenida, asumiéndose conocimientos en representación de conocimiento y ontologías, nociones de lenguajes de modelado de datos estándar como UML y arquitecturas de software. A continuación se describe la estructura de esta Tesis, organizada en seis capítulos y un apéndice.

Capítulo 1: Introducción. Este capítulo describe las motivaciones y las contribuciones de la Tesis.

Capítulo 2: Datos Enlazados y Ontologías. En este capítulo se incluye una descripción de la tecnología los Datos Enlazados y sus componentes principales (estándares de W3C).

Capítulo 3: Aplicaciones Especificas del Dominio. En este capítulo se incluye una descripción de conjuntos de datos enlazados y ontologías de referencia utilizados en Biodiversidad/Oceanografía, en base a esto se definen los requisitos que el sistema *BiGe-Onto* debe cumplir.

Capítulo 4: Desarrollo de *BiGe-Onto*. Se introduce el sistema *BiGe-Onto*, se presenta la metodología utilizada para el desarrollo y las partes que componen dicho sistema.

Capítulo 5: Evaluación de *BiGe-Onto*. En este capítulo se presenta la metodología utilizada para evaluar *BiGe-Onto*, y los casos de estudio utilizados para su validación. Además se compara *BiGe-Onto* con otras ontologías y sistemas similares.

Capítulo 6: Conclusiones. Se detallan los resultados y conclusiones de esta Tesis y se delinean líneas de investigación aún pendientes de estudio.

Apéndice: Preguntas de competencia. Se detalla como fueron implementadas las preguntas de competencia utilizadas en la validación.

Bibliografía: Recopilación de las referencias bibliográficas de todos los capítulos, ordenadas alfabéticamente por autor.

Capítulo 2

Datos Enlazados y Ontologías

Índice

En este capítulo, nos centramos en la definición del concepto de Datos Enlazados como una tecnología emergente con sus aplicaciones en diferentes áreas. Luego, describimos los principales componentes y estándares que hacen posible su funcionamiento.

2.1. Concepto de Datos Enlazados

La Web ha revolucionado la forma en que publicamos, accedemos y compartimos información. El uso de navegadores web permite navegar de forma transversal el espacio de información mediante enlaces de hipertexto. Los motores de búsqueda indexan los documentos web y analizan sus estructuras de enlaces para mejorar la búsqueda por cadenas a los usuarios. Esta funcionalidad ha contribuido al rápido crecimiento de la Web debido a su naturaleza flexible, abierta y extensible [BL06].

A pesar de todos los beneficios que la Web ha proporcionado desde que surgió, los enfoques tradicionales no son suficientes a nivel de datos. Los datos que se han publicado en la Web tienen diferentes formalismos. Sin embargo, la mayoría de estos datos carecen de significado y los enlaces que conectan documentos HTML no proporcionan información sobre la semántica de las relaciones entre entidades. Por lo tanto, es necesaria una tecnología más avanzada para conectar entidades con la suficiente expresividad. En este contexto, donde el espacio de información ha aumentado y los datos están vinculados entre sí, han surgido los Datos Enlazados (o en inglés Linked Data). Los Datos Enlazados se definen como un conjunto de prácticas recomendadas para compartir, exponer y conectar piezas de información, datos y conocimientos mediante el uso de identificadores de recursos uniformes (Uniform Resource Identifiers, URIs) y RDF. Esta tecnología, apoyada por W3C, permite la publicación e intercambio de información de manera interoperable y reutilizable. Linked Data tiene la propiedad de ser interpretada por una máquina, proporcionar un significado a los datos y posibilitar las interconexiones de otros conjuntos de datos externos a sí mismo, y puede vincularse desde y hacia conjuntos de datos externos [BLHL+01].

Los principios de Datos Enlazados se pueden resumir como Tim Berners-Lee propuso en [JHA+14]:

1. Usar URIs como nombres para las cosas.

2. Usar URIs que sean interpretables por humanos y máquinas.

3. Proveer información útil acerca de cada URI en algún estándar de la web (ej: RDF)

4. Crear enlaces entre URIs.

El primer principio se refiere al uso de las referencias URI para identificar no solo documentos web, sino también conceptos reales o cosas abstractas. El uso de URIs se utiliza para identificar recursos para conceptos abstractos o reales. El segundo principio se basa en el uso de URIs que en la Web tradicional recuperan información mediante un identificador único. Sin embargo, en Datos Enlazados, estos URIs usan el protocolo HTTP para recuperar una descripción de los conceptos abstractos o reales que representan, a este proceso se lo denomina *desreferenciar*. Para poder usar y recuperar los datos interconectados y crear aplicaciones que usen esta tecnología, los estándares son muy importantes. Esto ha llevado al tercer principio de los datos Enlazados que fomenta el uso de RDF, que es un marco para modelar información. La representación de la información RDF se caracteriza por estar mínimamente restringida y ser muy flexible, como lo especifica W3C en [W3C04]. Este marco se describe con más detalle en la Sección 2.2.1.

El cuarto principio se refiere al uso de enlaces RDF. Este nombre se usa para distinguir de los hipervínculos que simplemente conectan documentos web clásicos. Los enlaces RDF asociados a los Datos Enlazados no solo conectan cosas sino que también proporcionan información sobre el tipo de asociación entre las entidades que vinculan.

En resumen, estos principios en los que se basan los Datos Enlazados pueden considerarse como una receta para la comunidad cuyos esfuerzos principales se han centrado en 1) la publicación de conjuntos de datos de licencias abiertas como datos abiertos Enlazados en la web, 2) la interconexión de datos a otros repositorios de datos Enlazados y 3) desarrollar clientes para usar los Datos Enlazados.

Desde la propuesta de Datos Enlazados de Tim Berners-Lee en 2006, la Nube de Datos Abiertos Enlazados ha aumentado exponencialmente en los últimos años, creando lo que se llama el espacio de datos global de diferentes fuentes de datos. Los temas de publicación de Datos Enlazados son muy diferentes y van desde el gobierno, la geografía, las ciencias de la vida, la lingüística, etc. hasta los datos generados por los usuarios. En [HQD15], los autores realizaron una descripción general de las estadísticas de datos abiertos enlazados. Este estudio refleja cómo la información vinculada ha evolucionado a lo largo de estos años. Por ejemplo, en 2011, el número de conjuntos de datos fue de 452 y el número de tripletas RDF correspondió a 950 millones. En 2013, el número de conjuntos de datos y tripletas RDF aumentó hasta 2,289 y 11 billones, respectivamente. Actualmente, la nube de datos Enlazados está formada por 9960 conjuntos de datos y 149 billones de triples [LOD19]. El estado actual de la nube de datos Enlazados se muestra en la Figura 2.1.

2.2. Componentes de los Datos Enlazados

La Web ha facilitado la difusión de información en todo el mundo. La estructura web proporciona lo que se denomina URL (Uniform Resource Locator) para hacer referencia a las páginas web y conectarlas entre sí. Sin embargo, el concepto de Datos Enlazados se basa en el soporte de una Web de datos donde los datos están interconectados a través de URIs. El objetivo principal de los datos Enlazados es relacionar datos de diferentes fuentes y ser interpretables por máquinas. Los datos deben tener un significado, que se define explícitamente y se utilice para enlazar con otras fuentes de datos externas. El segundo y tercer principio de los datos Enlazados se refieren al uso de RDF como modelo de datos para representar esta información, que se describe en sección 2.2.1. La sección 2.2.2 detalla el lenguaje de consulta SPARQL utilizado para recuperar información de los puntos finales que actúan como pasarelas de consulta en relación a las fuentes de datos vinculadas semánticamente. En la sección 2.2.3, también hemos incluido el lenguaje

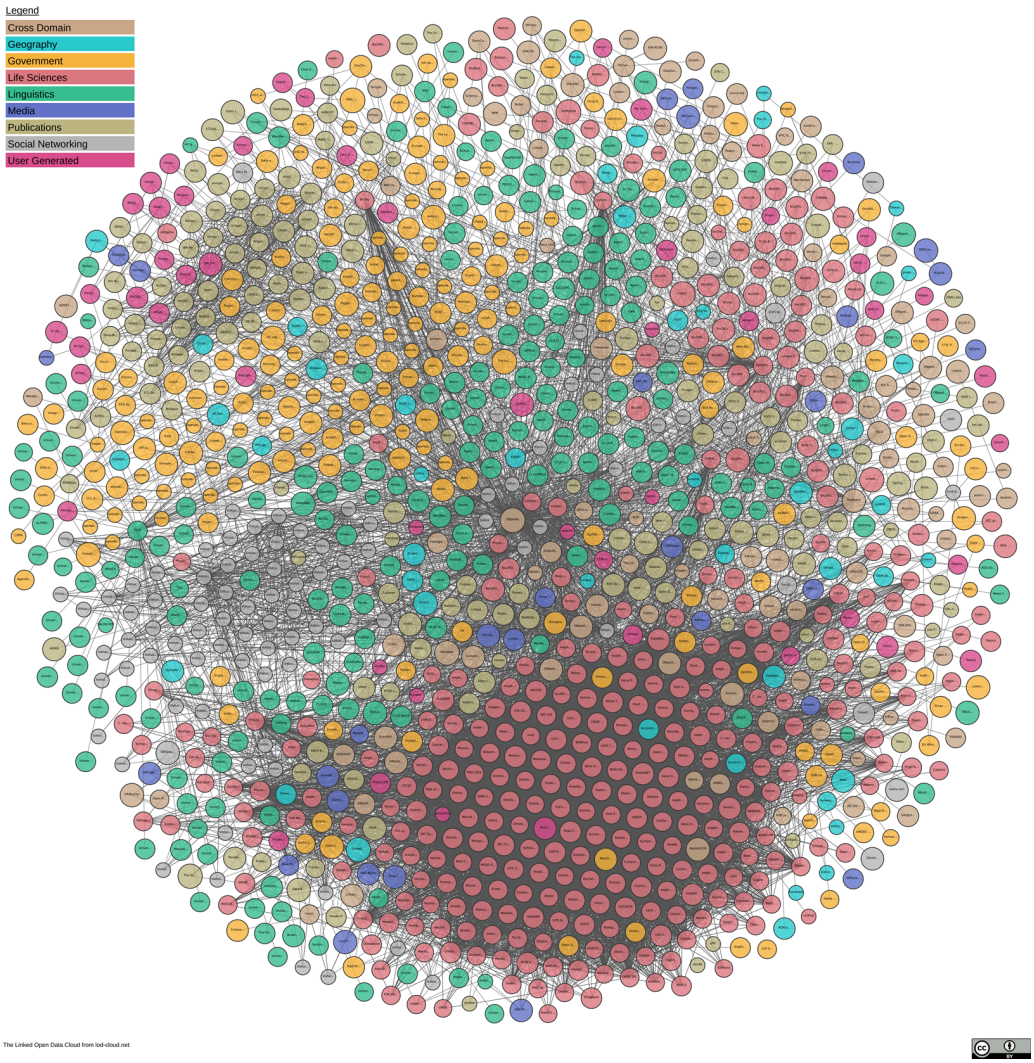

Figura 2.1: Estado actual de la nube de Datos Enlazados, extraído de la página web oficial Linked Open Data Cloud (2019) [LOD19]

OWL que extiende la expresividad RDF con primitivas adicionales y forma parte de los vocabularios de Datos Enlazados junto con RDF y el Sistema de Organización de Conocimiento Simple (SKOS [MB09] por sus siglas en inglés).

Después de que surgió el concepto de Datos Enlazados, se crearon conjuntos de datos RDF interconectados como se muestra en la Figura 2.1, creando la Web de Datos que incluye datos de dominios cruzados, Datos de Cien-

cias de la Vida, datos gubernamentales que se describen en detalle en la sección 2.2.4. En este contexto, se han realizado esfuerzos para investigar y crear aplicaciones para explotar Datos Enlazados como (ver sección 2.2.5) motores de búsqueda orientados a los humanos, navegadores, aplicaciones específicas de dominio, mash-ups de Datos Enlazados, etc. sección 3.1 describe la aplicación de la tecnología Linked Data en el dominio de las ciencias de la vida y el papel de las ontologías y terminologías en la Biodiversidad y Oceanografía.

2.2.1. Marco de Descripción de Recursos (RDF)

El Marco de Descripción de Recursos (del inglés Resource Description Framework) es un modelo de datos orientado a grafos para representar información en la Web soportada por el consorcio W3C [W3C04]. RDF representa la información en forma de grafos dirigidos etiquetados como nodos y arcos. Este modelo de datos fue diseñado para facilitar la integración de la información de fuentes de datos heterogéneas en diferentes modelos de datos. El objetivo era crear un estándar de modelo de datos que proporcionara interoperabilidad entre las aplicaciones que intercambian información comprensible por las máquinas en la Web. El diseño de RDF ha cumplido los siguientes objetivos 1) proporcionar una sintaxis basada en XML que sea la base de la tecnología web actual, 2) ofrecer un modelo de datos simple y abierto para representar información, 3) un lenguaje basado en URIs para representar recursos que puede vincularse a otros recursos y, finalmente, 4) la semántica (o significado) para representar el tipo de relaciones entre las cosas y con inferencia demostrable.

En RDF, un recurso se representa como un número de tripletas: *sujeto, predicado y objeto (nodo-arco-nodo)*. El sujeto de una tripleta es una URI que identifica el recurso que esta describiendo. En este caso, es un identificador único de una ocurrencia[1]. El objeto puede ser un literal (un número,

[1] En el contexto de la Biodiversidad, ocurrencia es la existencia de un organismo en un lugar particular en un momento particular.

una cadena de texto) u otro recurso que a su vez puede ser objeto de otra tripleta RDF. El predicado, enlaza el sujeto y el objeto en una tripleta RDF y representa su tipo de relación. También está representado por una URI. En este ejemplo el predicado representa que la ocurrencia 000faef2-bd8d-49c4-972e-e8a39777f54a fue registrada por Guillermo Suarez.

El sujeto de un tripleta RDF es una URI o un nodo en blanco (un URI que no es una referencia URI o literal o un URI sin un nombre intrínseco). El predicado es siempre una URI. El objeto puede ser una URI (que puede ser el sujeto de otra tripleta), un literal o un nodo en blanco.

Definición 1. Formalmente, una tripleta RDF se define como: Sea U el conjunto de URIs, L el conjunto de literales y B el conjunto de nodos en blanco. Una tripleta $(s,p,o) \in (U \cup B) \times U \times (U \cup L \cup B)$ se llama tripleta RDF.

Hay dos tipos de tripletas RDF: literales y enlaces RDF. Un literal se utiliza para describir una cadena de texto, números o fechas y puede clasificarse como simple o tipada. Simple se refiere a una cadena combinada con una etiqueta de idioma opcional. Tipada es cuando el literal se combina con una URI que identifica el tipo de datos del literal. Un ejemplo es: `<xsd:boolean, "true">` que hace referencia a que el literal especificado en la tripleta es un valor booleano y denota el verdadero lógico que se define en el esquema XML [XML08]. Un enlace RDF describe la relación entre dos recursos. Un conjunto de tales tripletas RDF conforma un grafo RDF, que se puede visualizar como un nodo y un diagrama de arco dirigido en el que cada tripleta se representa como un enlace *nodo-arco-nodo*. La Figura 2.2 muestra la tripleta que se refiere a la ocurrencia identificada de manera única con el UUID `<000faef2-bd8d-49c4-972e-e8a39777f54a>` (sujeto). En este ejemplo, el predicado proviene del vocabulario Darwin Core [WBG+12] (Ver Sección 3.2) y el objeto y el sujeto provienen del conjunto de datos *BiGe-Onto*.

Definición 2. Un grafo RDF G se puede definir como un conjunto de

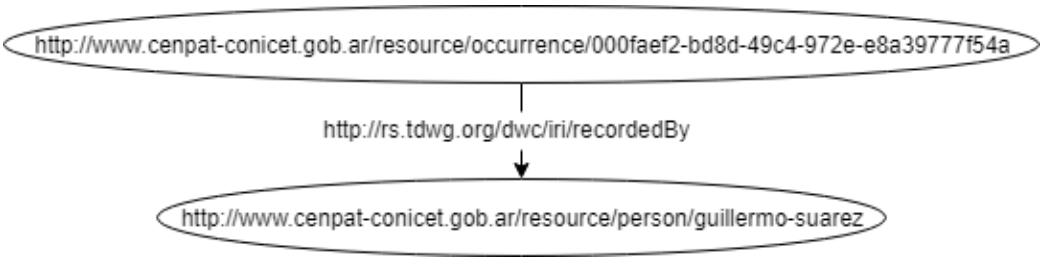

Figura 2.2: Representación gráfica de una tripleta RDF, el sujeto (ocurrencia) y el objeto (persona) es conectado por el predicado (recordedBy).

tripletas RDF. Entonces, *(s, p, o)* se puede representar como un grafo etiquetado con borde directo $s \xrightarrow{p} o$.

Vale la pena señalar que un grafo RDF no es un grafo clásico, dado que un predicado (nodo) puede aparecer como nodos de otros bordes como el caso de los gráficos RDF bipartitos [HG04]. Un ejemplo de un documento RDF que utiliza la sintaxis XML/RDF [RDF14] se muestra en la Figura 2.3, se muestra la representación de la ocurrencia mediante la URI `<occurrence/000faef2-bd8d-49c4-972e-e8a39777f54a>`

```
<?xml version="1.0" encoding="UTF-8"?>
 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <rdf:Description rdf:about="occurrence/000faef2....e8a39777f54a">
    <rdf:type rdf:resource="http://rs.tdwg.org/dwc/terms/Occurrence"/>
    <recordedBy xmlns="http://rs.tdwg.org/dwc/iri/"
    rdf:resource="person/guillermo-suarez"/>
    <label xmlns="http://www.w3.org/2000/01/rdf-schema#" xml:lang="en">
        000faef2-bd8d-49c4-972e-e8a39777f54a
    </label>
 </rdf:Description>
</rdf:RDF>
```

Figura 2.3: Serialización XML/RDF de la ocurrencia descripta anteriormente, por simplicidad algunas URIs fueron acortadas.

Un conjunto de datos RDF es una colección de grafos RDF y compromete un grafo predeterminado, que no tiene un nombre y puede estar vacío. También incluye cero o más grafos con nombre. Cada grafo nombrado es un par que consiste en una IRI o un nodo en blanco. Los nombres de los grafos

deben ser únicos.

Definición 3. Un conjunto de datos se puede denotar mediante (D,A), el cual es un grafo dirigido, donde D es el conjunto de nodos (que incluye URIs U, literales L y nodos en blanco B) y cada nodo $D_{\mathrm{i}} \in D$ denota un conjunto de datos; A es el conjunto de arcos y cada arco $(D_{\mathrm{i}},D_{\mathrm{j}}) \in A$ existe si hay al menos k tripletas $(s,p,o) \in D_{\mathrm{i}}$ donde s,o son dos URIs en $D_{\mathrm{i}},D_{\mathrm{j}}$ respectivamente. k indica los diferentes arcos en el grafo.

Después de la aparición de los datos Enlazados, el número de conjuntos de datos RDF disponibles ha aumentado. W3C proporciona una wiki donde están disponibles las URL de los volcados de datos RDF [Dat]. En esta página, los consumidores de Datos Enlazados pueden encontrar el nombre del proyecto, la URL del directorio que contiene los archivos de volcado de RDF e información sobre el editor o el mantenedor de los datos de RDF. Para consultar los datos RDF almacenados en los repositorios, los editores proporcionan un servicio llamado punto final (endpoint) donde los datos pueden consultarse con el lenguaje de consulta SPARQL [SPA08]. Un endpoint es una URL HTTP, que acepta consultas SPARQL y devuelve resultados. El servicio también puede devolver una variedad de serializaciones como Turtle [tur14], N-triples [ntr14], RDF/XML, etc. Las ventajas de que estos servicios funcionen se pueden resumir de la siguiente manera: 1) Acceso a los datos por parte de los consumidores de Datos Enlazados, 2) Ejecución de consultas complejas como consultas SPARQL federadas, que recuperan información de más de una fuente de datos, 3) Recuperación de la información que se representa con diferentes estándares para visualizar cómo se interconecta la información. En síntesis, podemos resumir un conjunto de ventajas definidas por el consorcio W3C de la siguiente manera:

- RDF se basa en lenguaje XML y admite cualquier tipo de datos XML.
- El modelo RDF compuesto por la estructura *sujeto-predicado-objeto* permite implementar y almacenar de manera eficiente los Datos Enlazados.

- RDF proporciona un vocabulario extensible de URIs.
- El modelo RDF permite hacer declaraciones sobre cualquier recurso.
- El modelo RDF se basa en grafos con etiquetas de borde directo, por lo que tiene la ventaja de estructurar la información mediante grafos.
- El modelo RDF se puede procesar en ausencia de información más detallada sobre la semántica. Por ejemplo, algunas inferencias se pueden encontrar lógicamente.
- La información almacenada en RDF tiene ventajas sobre las bases de datos relacionales basadas en: 1) En RDF, el esquema de datos es opcional, 2) La adición de nuevos atributos a un repositorio RDF se puede realizar sobre la marcha (en las bases de datos requiere migraciones), 3) En RDF, se pueden realizar inferencias, 4) RDF permite la detección de redundancias en el caso de incluir datos externos. 4) La estructura de las consultas permite uniones más rápidas.

2.2.2. SPARQL Lenguaje de consulta para RDF

SPARQL (protocolo SPARQL y lenguaje de consulta RDF) es un lenguaje de consulta RDF propuesto por el consorcio W3C para recuperar información de conjuntos de datos RDF y manipularla. La versión actual del lenguaje de consulta SPARQL es 1.1 [SPA13] y su lanzamiento oficial fue en 2013. La versión anterior de SPARQL era la 1.0 y se propuso oficialmente en 2008 [SPA08]. Como los autores describen en [AGP10], las consultas SPARQL están compuestas por tres partes: 1) la parte del patrón coincidente, que incluye varias características de la coincidencia de patrón del grafo, como las partes opcionales, la unión de patrones, el anidamiento, el filtrado de valores de posibles coincidencias y también la posibilidad de elegir la fuente de datos para que coincida con un determinado patrón; 2) los modificadores del resultado, que una vez que se ha calculado el patrón de coincidencia, permiten modificar los valores aplicando operadores clásicos como proyección, distinto, orden y

límites; 3) el resultado de la consulta SPARQL, que puede ser de diferentes tipos, como consultas sí/no, la construcción de un grafo a partir de estos valores como un conjunto de declaraciones RDF (sujeto-predicado-objeto). Una consulta SPARQL está estructurada de manera similar a una consulta SQL[2] con algunas diferencias. Una consulta SPARQL típica (Figura 2.4) incluye las siguientes partes, aunque algunas de ellas son opcionales:

- Las declaraciones de prefijos (`PREFIX`) que permiten abreviar URIs. Por ejemplo, la URI `<http://www.w3.org/1999/02/22-rdf-syntax-ns#>` puede ser abreviada como `rdf`.

- Una clausula `SELECT` que incluye todas las variables a proyectar en la consulta. Esta cláusula presenta algunas diferencias con respecto al SELECT de SQL. La cláusula `SELECT` en SPARQL se puede reemplazar por la cláusula `CONSTRUCT` que devuelve un grafo RDF especificado por una patrón, la cláusula `ASK` comprueba si un patrón de consulta tiene una solución y la cláusula `DESCRIBE` devuelve como resultado un único grafo RDF con datos sobre un recurso determinado.

- La cláusula `WHERE` proporciona el patrón de grafo básico para la concordancia con el grafo de datos. Este patrón incluye un conjunto de tripletas que pueden incluir variables y operadores. Estos patrones de grafos son los filtros para los valores que se devolverán.

- Una cláusula `FROM` que especifica que grafo se está consultando. Esta cláusula es opcional. Si la cláusula no se especifica en la consulta, la consulta SPARQL recupera información de todos los grafos dentro del conjunto de datos RDF.

- Finalmente, los modificadores de la consulta, que ordenan, dividen u organizan los resultados obtenidos de la consulta SPARQL. Por ejemplo, el caso de `ORDER BY` que establece el orden de la secuencia de resultados; `LIMIT` pone un límite máximo al número de tripletas que se

[2]Lenguaje de consulta estructurado para consultar bases de datos relacionales

devuelven; La cláusula `OFFSET` controla donde comienzan las tripletas dentro del conjunto global de tripletas que son devueltas.

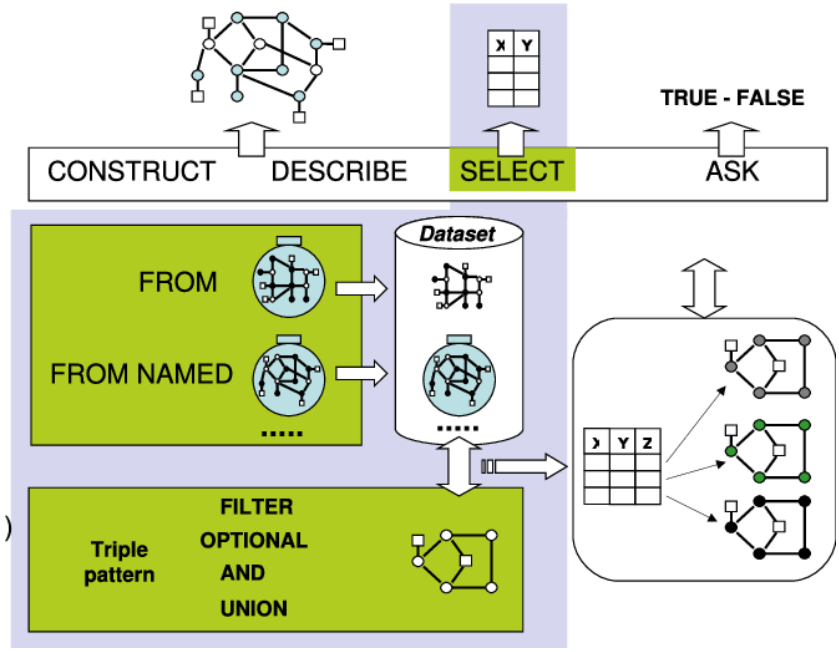

Figura 2.4: Forma general de una consulta SPARQL. La consulta puede incluir `SELECT`, `CONSTRUC`, `DESCRIBE` o `ASK`. La cláusula del conjunto de datos `FROM` o `FROM NAMED` especifica la URI o el nombre de un grafo especifico a ser consultado. La cláusula `WHERE` que proporciona el patrón RDF (puede incluir `FILTER`, `OPTIONAL` y `UNION`)

La sintaxis de SPARQL se basa en IRIs que extienden la sintaxis de las URIs a un abanico más amplio de caracteres. Por lo tanto, en SPARQL, los IRIs se utilizan para diseñar recursos. La sintaxis para literales es una cadena encerradas entre comillas dobles o simples y una etiqueta de idioma opcional, IRI o un prefijo. Las variables son declaradas anteponiendo el carácter "`?`" o "`$`", los cuales no son parte de la variable. Los nodos en blanco se designan mediante una etiqueta de la forma: "`_:abc`" o "`[]`". Los patrones de las tripletas se escriben como una lista separada por espacios en blanco entre sujeto, predicado y objeto.

La Figura 2.5 ilustra una consulta SPARQL. La primera línea corresponde con la cláusula `SELECT`, donde las variables `?occurrence` y `?name` fueron

declaradas incluyendo el modificador `DISTINCT`, esto asegura que se eliminen los resultados duplicados. Por lo tanto, el resultado de la consulta SPARQL devolverá una tabla con dos columnas con las variables asignadas (occurrence y name). La cláusula `WHERE` está compuesta por un patrón de tres tripletas: el primero (línea 8) significa que la consulta está buscando todos los sujetos de tipo `dwc:Occurrence`. El segundo patrón (línea 9) busca las personas `?person` que registraron las ocurrencias y finalmente el tercer patrón (línea 10) recupera los nombres de las personas `foaf:name`. También se pueden agregar filtros adicionales `REGEX` para restringir los resultados por nombre de las personas, pero en este caso no se especifica en la consulta SPARQL. Finalmente, el modificador `ORDER BY` (línea 12) ordena por nombre de las personas de manera ascendente.

```
1  PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
2  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
3  PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
4  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
5
6  SELECT DISTINCT ?occurrence ?name
7  WHERE {
8     ?occurrence rdf:type dwc:Occurrence.
9     ?occurrence dwciri:recordedBy ?person.
10    ?person foaf:name ?name
11 }
12 ORDER BY ?name
```

Figura 2.5: Ejemplo de consulta SPARQL.

Las consultas SPARQL se ejecutan contra conjuntos de datos RDF (consulte la **Definición 3**). Los desarrolladores y editores ofrecen un servicio SPARQL que acepta consultas y devuelve los resultados de las consultas a través de HTTP. Estos resultados pueden representarse en una variedad de formatos, como XML (que devuelve los resultados como tablas), un objeto JSON que es muy útil para aplicaciones web, RDF (resultados devueltos como declaraciones de sujeto-predicado-objetos), N-triples, turtle y HTML.

En este contexto, y con el surgimiento de la tecnología de Datos Enlazados, la cantidad de servicios que proporcionan información en los conjuntos de datos RDF que se deben consultar ha aumentado considerablemente,

brindando a los consumidores de datos la oportunidad de fusionar los datos distribuidos en la Web. Esto ha provocado la creación de una nueva especificación que define la semántica y la sintaxis de la extensión `SERVICE` en la versión 1.1 de SPARQL. Las consultas SPARQL que recuperan información de más de una fuente se denominan consultas federadas [QL08]. Estas consultas SPARQL contienen la palabra clave `SERVICE`, que indica a un procesador de consultas federadas que invoque una parte de una consulta SPARQL contra un conjunto de datos remoto. La consulta de la Figura 2.6 representa un ejemplo de una consulta federada SPARQL, la que recupera información de más de una fuente (el conjunto de datos *BiGe-Onto* y el conjunto de datos denominado DBPedia [DBP19]). La consulta está compuesta por varias partes: 1) Todos los prefijos; 2) la cláusula `SELECT` que declara todas las variables; 3) El patrón de las declaraciones sujeto-objeto-predicado; 4) La palabra clave `SERVICE` que invoca el servicio SPARQL de DBPedia y contiene un patrón RDF. La consulta del ejemplo hace la siguiente pregunta: *Obtener los estados de conservación de cada especie*

```
1  PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
2  PREFIX owl: <http://www.w3.org/2002/07/owl#>
3  PREFIX dbo: <http://dbpedia.org/ontology/>
4
5  SELECT ?scname ?c_status
6  WHERE {
7    ?s a dwc:Taxon.
8    ?s dwc:scientificName ?scname.
9    ?s owl:sameAs ?resource .
10   SERVICE <http://dbpedia.org/sparql> {
11      ?resource dbo:conservationStatus ?c_status.
12   }
12 }
```

Figura 2.6: Ejemplo de consulta federada. Esta consulta recupera información del conjunto de datos BiGe-Onto y de DBPedia. Esta consulta tiene la cláusula `SERVICE` que invoca un segundo servicio para obtener información sobre el estado de conservación de una determinada especie.

2.2.3. Lenguaje de Ontologías Web (OWL)

OWL [OWL09] es un lenguaje de representación de conocimiento estándar para la Web Semántica creado y recomendado por el consorcio W3C. El término representación del conocimiento se refiere al método de modelar el mundo real mediante el uso de entidades que representan cosas y relaciones entre estas entidades. OWL es un lenguaje de representación de conocimiento muy expresivo y flexible que ha tenido aplicaciones en varios dominios como atención médica, tráfico y automóviles, etc. OWL se basa en lógicas descriptivas (DL) [Baa09], lo que permite que este lenguaje se pueda razonar. OWL también se basa en la asunción de mundo abierto, lo que significa que lo que no se conoce como verdadero, es simplemente desconocido y no es necesariamente falso. La versión actual de OWL se llama OWL 2 [OWL12] que es una extensión de OWL y permite más expresividad que su predecesora. En [GHM+08], los autores analizan las deficiencias identificadas de OWL 1, como problemas de expresividad, problemas con sus sintaxis y definiciones de los diferentes subconjuntos de OWL. Además, los autores presentan una descripción general de OWL 2 y cómo esta nueva versión desarrollada por W3C supera los problemas presentados en OWL 1. En este capítulo, presentamos una descripción general de la versión de OWL 2 tal como la hemos utilizado. El elemento más importante en una ontología OWL 2 son los IRIs, que representan una entidad del mundo real. Como se especifica en [OWL12], estos IRIs representan una ontología y sus elementos son absolutos y no relativos. En una ontología dada, dos IRIs son estructuralmente equivalentes si la construcción de la cadena es idéntica. Los IRIs se encierran entre caracteres $<$ y $>$. Los IRIs pueden ser muy largos, por lo tanto, estos se abrevian como en las consultas SPARQL con un nombre de prefijo *pn:* seguido de una cadena vacía, el carácter : (dos puntos) y a continuación asociarse con un prefijo IRI. Los prefijos se pueden especificar al principio del documento OWL y se pueden leer mediante analizadores OWL. Las clases, propiedades (relaciones entre clases), individuos y tipos de datos son las construcciones básicas de una ontología OWL. Las clases representan conceptos de un dominio, por ejemplo, ocurrencias, eventos, personas, buques, etc.

En OWL 2, hay clases con la IRI *owl:Thing* que representa el conjunto con todos los individuos y *owl:Nothing* que corresponda a un conjunto vacío. Las instancias de las clases se llaman individuos, por ejemplo, dada una clase *owl:Region* cuyos limites geográficos se han modelado en OWL, y si un punto que se encuentra dentro de esos limites puede considerarse como un individuo de *Region*. Las propiedades representan las relaciones entre las clases. Hay dos tipos de propiedades en OWL 2: 1) propiedades de datos (*OWLDataProperty*) que representan la relación entre una clase y un tipo de datos que puede ser una cadena, un entero, un booleano, etc. 2) propiedades de objeto *OWLObjectProperty* que relacionan dos clases. La Figura 2.7 muestra un fragmento de una ontología OWL basada en el dominio de la Biodiversidad:

```
1 <owl:Class rdf:about="http://purl.org/dc/terms/Location">
2   <rdfs:subClassOf rdf:resource="bigeonto:Region"/>
3   <rdfs:comment xml:lang="en">
2        A spatial region or named place.
3    </rdfs:comment>
4   <rdfs:isDefinedBy xml:lang="en">
5        http://purl.org/dc/terms/
6    </rdfs:isDefinedBy>
7   <rdfs:label xml:lang="en">Location</rdfs:label>
8 </owl:Class>
9
10 <owl:ObjectProperty rdf:about="http://rs.tdwg.org/dwc/iri/recordedBy">
11  <rdfs:domain rdf:resource="http://rs.tdwg.org/dwc/terms/dwc:Occurrence"/>
12   <rdfs:range rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
13   <rdfs:comment xml:lang="en">Provides information about persons,
14   who have recorded an occurrence. It is also reused from the DwC URI
15   namespace and enables non-literal ranges for its analoguos with DwC
16   http://tdwg.github.io/dwc/terms/index.htm#recordedBy .
17   </rdfs:comment>
18   <rdfs:isDefinedBy rdf:resource="http://rs.tdwg.org/dwc/iri/"/>
19   <rdfs:label xml:lang="en">recorded by</rdfs:label>
20 </owl:ObjectProperty>
21
22 <owl:DatatypeProperty rdf:about="dwc:scientificName">
23   <rdfs:domain rdf:resource="http://rs.tdwg.org/dwc/terms/dwc:Taxon"/>
24   <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
25 </owl:DatatypeProperty>
```

Figura 2.7: Fragmento de ontología definida en OWL.

La línea 1 define la clase denominada ubicación (`Location`), la línea 2 define que `Location` es una subclase de `Region`, de la línea 3 a la 5 se utilizan

comentarios y etiquetas para proporcionar información de la clase `Location`. En las líneas 8 a 18, se define una *ObjectProperty* llamada `RecordedBy` que relaciona la clase `Occurrence` con la clase `Person`. Por último (líneas 20 a 23) se muestra la definición de una *DatatypeProperty* llamada `scientificName`.

Existen diversos dominios como la Biodiversidad, donde la taxonomía es un requisito esencial para modelar cómo se organizan los datos. Una jerarquía taxonómica puede modelarse utilizando clases, subclases, axiomas separados, axiomas equivalentes, propiedades de objetos y datos, etc. En OWL 2 hay construcciones que definen clases más complejas, y que son muy útiles para dar mayor expresividad al describir clases que comparten algunas propiedades. Algunas de estas construcciones son:

1. *owl:intersectionOf* que se define como $C \equiv A \cap B$ esto significa que los individuos en C también son miembros de A y B

2. *owl:unionOf* que se define como $C \equiv A \cup B$ esto significa que los individuos en C también son miembros de A y B o de ambos.

3. *owl:complementOf* que se define como $A^{\mathrm{C}} = \{x \in U \mid x \notin A\}$ esto significa que si U es el universo que contiene todos los elementos y x no está contenido en A, entonces x está contenido en A^{C}.

Además de estas expresiones de clase complejas, existen otras que se pueden usar como restricciones de propiedad. Estas restricciones de propiedad se utilizan para imponer restricciones a una clase de OWL determinada. Estas restricciones de propiedad son *owl:allValuesFrom*, *owl:someValuesFrom*, *owl:maxQualifiedCardinality*, *owl:minQualifiedCardinality*, etc. Un ejemplo de esto se aprecia en la Figura 2.8:

El fragmento OWL anterior representa la clase `Occurrence`, que tiene como propiedad objeto `recordedBy` y define una restricción sobre la misma (*owl:maxQualifiedCardinality*) con valor igual a 1. Esto significa que para cada ocurrencia existirá como máximo una persona que la registre.

```
1<owl:Class rdf:about="http://rs.tdwg.org/dwc/terms/dwc:Occurrence">
2 <rdfs:subClassOf>
3 <owl:Restriction>
4  <owl:onProperty rdf:resource="http://rs.tdwg.org/dwc/iri/recordedBy"/>
5  <owl:maxQualifiedCardinality
6  rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger"> 1
7  </owl:maxQualifiedCardinality>
8  <owl:onClass rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
9 </owl:Restriction>
10</rdfs:subClassOf>
11</owl:Class>
```

Figura 2.8: Restricción de cardinalidad definida en OWL

Las propiedades en las ontologías juegan un papel importante, ya que describen relaciones entre clases OWL. Las propiedades se relacionan entre sí mediante términos como: *rdfs:subPropertyOf*, *rdfs:equivalentProperty*, etc. Semánticamente, esto significa que, si una propiedad llamada p_2 es una subpropiedad de otra propiedad llamada p_1, todas las instancias que están relacionadas por la propiedad p_2 también está relacionadas por la propiedad p_1. OWL 1 y OWL 2 no permiten constructores como conjunción, disyunción, etc. Sin embargo, OWL 2 proporciona construcciones, que son cadenas de propiedades que son muy útiles en aquellas ontologías en las que es necesario modelar propiedades complejas. Las propiedades de OWL tienen características diferentes, como dominios, rangos, características reflexivas, transitivas y asimétricas, etc.

```
1  <owl:Class rdf:about="http://purl.org/dc/terms/Location">
2   <rdfs:subClassOf rdf:resource="bige-onto/Region"/>
3   <rdfs:comment xml:lang="en">
4      A spatial region or named place.
5   </rdfs:comment>
6   <rdfs:isDefinedBy xml:lang="en">
7   http://purl.org/dc/terms/
8   </rdfs:isDefinedBy>
9   <rdfs:label xml:lang="en">Location</rdfs:label>
10 </owl:Class>
```

El ejemplo anterior muestra la clase `Location`, la cual es subclase de `Region`. Esto significa que todas las instancias incluidas en la clase `Location` pertenecen a la clase `Region`.

Las ontologías OWL 2 pueden aprovechar los razonadores para infe-

rir el conocimiento de ellas. Los razonadores son una pieza de software que se ha desarrollado en los últimos años, algunos de los mas conocidos son: Hermit [GHM+14], Pellet [SPG+07], CEL [SPG+07], TrOWL [TPR10], Elk [KKS14]. Todos estos razonadores pueden admitir herramientas de visualización como Protégé [Mus15] o NeOn toolkit [HLS+08]. Estos razonadores presentan varias características, que se pueden enumerar de la siguiente manera. La satisfacibilidad de la ontología consiste en saber si la ontología es consistente o no. Si una ontología dada presenta una inconsistencia, se refiere al hecho de que la ontología tiene una contradicción y, por lo tanto, dos afirmaciones contradictorias. Por ejemplo, dada una instancia a (definida en una ontología O) a está clasificada por el razonador en dos clases A y B modeladas como clases disjuntas. La verificación de instancias chequea si una instancia llamada a está clasificada en la clase A. La satisfacibilidad de clase se refiere a si una clase A puede tener instancias. La subsunción, verifica si dos clases llamadas C y D, $C \sqsubseteq D$ entonces C es una subclase de D. La clasificación genera todas las relaciones de subclases. OWL 2 presenta diferentes subconjuntos (también llamados perfiles) basados en la expresividad y los requisitos de escalabilidad. Cuanto más expresivo sea el perfil OWL 2, mayor será la complejidad. El perfil *OWL DL* es el más expresivo y permite conservar la integridad, la decidibilidad y la disponibilidad de algoritmos prácticos de razonamiento. *OWL Lite* esta pensado para los usuarios que necesitan una clasificación jerárquica y restricciones simples como valores de cardinalidad entre 0 y 1. *OWL full* se basa en la semántica de *OWL Lite* y *OWL DL*, con cierta compatibilidad con el esquema RDF y sin restricciones. *OWL full* es indecidible, lo que significa que ningún razonador puede completar el razonamiento. En *OWL EL* la complejidad de las tareas de inferencia es polinomial. Este perfil puede ser muy útil para aplicaciones que necesitan un gran número de clases y propiedades, pero que no requieren construcciones complejas de OWL. *OWL QL* está diseñado para admitir consultas conjuntas en bases de datos relacionales. Este perfil tiene la peor complejidad en términos de tiempo polinomial. *OWL RL* permite a un sistema basado en reglas realizar un razonamiento en tiempo polinomial.

2.2.4. La Web de los Datos

En las secciones anteriores, hemos descrito los lenguajes estándares para representar y recuperar información (RDF, SPARQL y OWL) utilizados en la tecnología de datos Enlazados. En esta sección, describimos la Web de datos y sus propiedades, la topología de la misma y los datos Enlazados generados en diferentes dominios.

El concepto de Web de los Datos (en inglés *Web of Data*) se define como un espacio global de datos generado por un gran número de individuos e instituciones (privadas y públicas) que publican sus datos. Este espacio de datos resultante incluye todo tipo de información referente a geografía, publicaciones científicas, gobiernos, música, programas de radio y televisión, ciencias de la vida, etc. La Web de datos presenta un conjunto de propiedades resumidas en [HB11] que se pueden enumerar aquí de la siguiente manera:

- La Web de los Datos es genérica y contiene todo tipo de información.
- Todas las instituciones e individuos son libres de publicar datos en este espacio de datos Enlazados.
- Este espacio global puede incluir desacuerdos o información contradictoria entre entidades.
- Todas las entidades están conectadas a través de enlaces RDF. Toda la información representada como RDF crea un grafo que abarca otras fuentes permitiendo el descubrimiento de nuevas fuentes de datos. De acuerdo a esto, las aplicaciones que consumen datos Enlazados pueden detectar nuevas fuentes agregadas en tiempo de ejecución al seguir los enlaces RDF que conectan entidades.
- Los publicadores de datos no tienen restricciones en el uso de un vocabulario específico para representar sus datos.
- Los datos se autodescriben, lo que significa que si una aplicación consume un conjunto de datos que siguen un vocabulario no especificado, la

aplicación debe poder desreferenciar los URI que identifican los términos que representan los conceptos.

- El uso del protocolo HTTP como un vocabulario estándar siguiendo los principios de datos Enlazados.

La Web de los datos es soportada por la comunidad de la Web Semántica y el proyecto Linked Open Data. Este proyecto tiene como objetivo principal detectar fuentes de datos bajo licencias de código abierto, convertir los datos en tripletas RDF y publicar los datos en la Web. El principio fundamental de esta comunidad es que cualquier usuario puede contribuir al aumento de la Web de datos. Esta apertura que caracteriza a este proyecto puede considerarse como un factor importante en términos de éxito. Esto puede ser confirmado empíricamente por el rápido crecimiento que la Web de los datos ha experimentado en los últimos años de acuerdo con las últimas estadísticas proporcionadas en [LOD19].

Según [PPMT18], la nube de datos abiertos enlazados entre Agosto de 2014 y Mayo de 2019 contenía los siguientes conjuntos de datos, distribuidos en los siguientes dominios: gobierno (23.85 %), publicaciones (23.33 %), red social (15.78 %), ciencias de la vida (11.05 %), dominio cruzado (7.19 %), contenido generado por el usuario (7.36 %), Geográfico (4,21 %), medios de comunicación (3,68 %), lingüística (3,50 %). Los gobiernos y las instituciones públicas generan datos valiosos, que se almacenan y con frecuencia no se utilizan. Los datos gubernamentales incluyen información que abarca desde la economía hasta las estadísticas de los ciudadanos, los informes de las instituciones públicas, etc. Toda esta información se ha incrementado en la última década, especialmente con la tendencia de abrir los datos gubernamentales al público. Por lo tanto, en este contexto, hay algunas iniciativas, como data.gov.uk y data.gov, que vinculan su información con otros conjuntos de datos utilizando RDF. Los datos de RDF han sido publicados por estos proyectos siguiendo algunas pautas sobre cómo publicar datos gubernamentales propuestos por [Tim09, VTVBCGP11]. Dentro de la categoría de publicaciones, existen conjuntos de datos con información sobre bibliotecas,

publicaciones científicas, bases de datos de citas, etc. Ejemplos de proyectos que siguen los principios de Datos Enlazados son el Catálogo de la Unión Sueca llamado LIBRIS[3], que comenzó a compartir datos Enlazados en 2008. Los datos Enlazados relacionados con la categoría generada por el usuario son generados por portales que recopilan información de las comunidades de los usuarios. Algunos ejemplos son datos sobre blogposts publicados como Linked Data por wordpress.com, datos de software de código abierto como apache.org, flujos de trabajo en de experimentos cientificos publicados en myExperiment.org y reseñas generadas en goodreads.com y revyu.com.

La categoría de red social incluye datos RDF de los perfiles de los usuarios sociales y las conexiones entre estas personas, incluidos los perfiles FOAF y los vínculos generados por StatusNet. La categoría generada por el usuario es diferente de la categoría de red social ya que la primera contiene información sobre los datos publicados por los usuarios en páginas que involucran a una comunidad y la última sobre los perfiles sociales de los usuarios públicos.

Los datos de dominios cruzados, que es otra categoría de la nube de datos Enlazados, se refieren a información que no es específica de un tema determinado. Este tipo de dominio para datos es muy necesario para conectar conjuntos de datos específicos a otros, evitando fragmentos no interconectados de la nube de datos Enlazados en términos de temas. Un ejemplo de estos datos Enlazados de dominio cruzado es DBpedia [DBP19], que es un proyecto que extrae información estructurada de Wikipedia[4] y hace que esta información esté disponible en la Web de datos. Desde el inicio de este proyecto, DBpedia ha servido como eje central para el surgimiento de los Datos Enlazados. Además, en las últimas décadas, el enfoque principal de DBpedia ha sido la extracción de información de artículos de Wikipedia de infoboxes, imágenes, categorización de la información, citas, enlaces, etc. Otros conjuntos de datos de dominios cruzados adicionales incluyen recursos lingüísticos como WordNet.

[3] http://librisbloggen.kb.se/2008/12/03/libris-available-as-linked-data/
[4] https://en.wikipedia.org/wiki/Main_Page

Los conjuntos de datos en la categoría geográficos se relacionan con varios dominios que tienen antecedentes sobre la geografía. Por ejemplo, GeoNames que fue el primer conjunto de datos que ofreció información geográfica como Datos Enlazados y que sirve como un eje central en este tema. Otro conjunto de datos significativo, que forma parte de los datos geográficos vinculados es LinkedGeoData. Este conjunto de datos RDF recopila información de OpenStreetMap (una fuente de datos espaciales) y la pone a disposición de los servicios REST. DBpedia también proporciona información RDF geoespacial y está interconectada con GeoNames y LinkedGeoData, lo que contribuye al aumento de la nube de datos Enlazados [SLHA12]. La categoría de medios de comunicación contiene conjuntos de datos que proporcionan información sobre música, películas, TV y radio. Uno de los conjuntos de datos más destacados dentro de esta categoría es la British Broadcasting Corporation (BBC), que publica grandes cantidades de contenido en la Web, como textos, videos y audio. Sin embargo, la mayor parte de esta información es accesible a través de micrositios HTML específicos, excluyendo una integración de datos más amplia con el resto de la información de la BBC. Este contexto involucra problemas relacionados con la vinculación entre dominios cruzados y la desambiguación entre vocabularios. Por lo tanto, para abordar estos problemas, los autores en proponen en [KSR+09] utilizar DBpedia, que proporciona un vocabulario estándar para vincularse con los datos de la BBC siguiendo los principios de Datos Enlazados. El trabajo resultante condujo a un sistema de categorización llamado CIS para interconectar elementos de datos y un conjunto de servicios que utilizan tecnología vinculada subyacente. Del mismo modo, otro conjunto de datos importante dentro de esta categoría que sigue el mismo enfoque es The New York Times, que publica datos como RDF al interconectar datos con conjuntos de datos existentes, como DBpedia, GeoNames, etc.

La categoría ciencias de la vida se han convertido en una rama muy importante que ha estado generando cantidades de información a medida que las tecnologías han avanzado en la era posgenómica. Toda esta información contenida en bases de datos aisladas se ha convertido a RDF de acuerdo con

los principios de Datos Enlazados, contribuyendo a la Web de Datos. Ejemplos de proyectos destacados en los datos Enlazados de ciencias de la vida son: Bio2RDF [BNT+08], un sistema que ayuda a resolver el problema de la integración del conocimiento en el dominio de las ciencias de la vida. El proyecto Bio2RDF se puede ver como un integración de categorías, ya que combina información de 35 conjuntos de datos de Life Sciences diferentes. El número de tripletas RDF que ha sido generado por Bio2RDF corresponde a 11 millones. Otro proyecto es BioPortal [NSW+09], un repositorio abierto de ontologías biomédicas que almacena más de 700 ontologías desarrolladas en diferentes formatos como OWL, OBO , RRF. Ambos proyectos han contribuido a enriquecer el espacio web en la categoría Ciencias de la Vida. Todas estas contribuciones han sido realizadas por una comunidad de usuarios que comprende investigadores, organizaciones, etc., lo que permite la evolución del espacio de datos al estado actual.

2.2.5. Aplicaciones de Datos Enlazados

Tras la aparición de la tecnología Linked Data, las llamadas aplicaciones de Linked Data han tenido un papel importante. De acuerdo con [Hau09], hay dos formas de definir el término *aplicaciones de datos Enlazados*; el primero se refiere a las aplicaciones de Datos Enlazados en dominios específicos como ciencias de la vida, medios de comunicación, etc. y el segundo corresponde a las aplicaciones que se implementan en la capa superior de los Datos Enlazados. Ambos tipos de interpretaciones sobre aplicaciones de datos Enlazados no se excluyen entre sí y deben ser complementarios. Por lo tanto, podemos definir las aplicaciones de Datos Enlazados como aquellas que están diseñadas para consumir y manipular datos Enlazados. Estas aplicaciones de datos Enlazados se pueden dividir en dos grupos: 1) aplicaciones genéricas y 2) aplicaciones específicas de dominio. El primer grupo de aplicaciones puede procesar información de cualquier dominio y el segundo se refiere a las aplicaciones que consumen datos vinculados de un dominio específico. Entre las aplicaciones genéricas de datos vinculados, también hay

dos categorías: los navegadores de datos vinculados y los motores de datos vinculados. Los navegadores para Datos Enlazados fueron uno de los primeros sistemas implementados que aprovecharon esta tecnología. De manera similar a los tradicionales, estos navegadores navegan a través de la Web utilizando enlaces RDF y representan los recursos RDF y sus propiedades. Por ejemplo, Haystack es una aplicación implementada en 2004 que agrega RDF de múltiples ubicaciones, lo que permite a los usuarios navegar a través de diferentes grafos RDF. Este navegador también proporciona un modelo de colecciones para recopilar y estructurar información. Disco Hyperdata [DIS07] permite a los usuarios representar información RDF como páginas HTML. Noadster [RVOH05] proporciona una interfaz de documento hipermedia para información codificada en semántico Estándares web Esta aplicación proporciona una interfaz de solicitud de información que devuelve una lista de puntos de inicio para navegar. Piggy Bank [HMK05] es una extensión de navegador web que reestructura la información de las páginas HTML en RDF para generar información en formato Web Semántico para que la consuman las aplicaciones. Tabulator [BLCC+06] un navegador RDF diseñado para que los usuarios finales interactúen y naveguen a través de los recursos RDF y para los desarrolladores de contenido RDF como incentivo para ellos al publicar y refinar su información en RDF y mostrar cómo este nuevo contenido RDF interactúa con otra información. LENA [KFS08] proporciona diferentes vistas de los datos, siguiendo los criterios del usuario que se expresan como consultas SPARQL. Visor [PSHS11] es un navegador basado en múltiples pivotes, que se puede configurar en cualquier punto final SPARQL. Esta herramienta permite a los usuarios explorar clases y propiedades específicas del esquema RDF de las colecciones y crear hojas de cálculo con la información seleccionada. Otros navegadores de facetas son Humboldt [KD08], facet [HVOH06] y gFacet [HZL08]. Explorator [ASB09] es una herramienta de búsqueda exploratoria de código abierto para información RDF, que implementa una interfaz de manipulación directa. Implementa un conjunto de operaciones personalizadas y algunos ejemplos de consultas SPARQL. Information Workbenches [HHSS12] otra herramienta de exploración para RDF que ofrece varias herramientas de back-end y front-end.

Marbles [Mar09] es un proyecto que utiliza el vocabulario de Fresnel para representar información RDF como páginas HTML. Como la cantidad de información RDF ha aumentado con el surgimiento de la tecnología de Datos Vinculados, buscar datos y proporcionar información útil en la solicitud de cualquier usuario es un desafío. Varios motores de búsqueda, categorizados como aplicaciones genéricas de datos vinculados, rastrean la Web de datos y proporcionan información agregada. Ejemplos de estos esfuerzos son Swoogle [DFJ+04], que proporciona búsquedas en documentos RDF mediante un índice de palabras clave invertidas y una base de datos relacional. Swoogle también tiene algunas métricas para calcular la popularidad de las clases y las relaciones. Watson otro navegador semántico, que es muy similar a Swoogle y proporciona una búsqueda basada en palabras clave para encontrar documentos y un servicio de API. Sindice [TDO07] ofrece un servicio de búsqueda basado en Lucene y MapReduce. El navegador Falcons [CGQ08] proporciona una búsqueda centrada en la entidad de conceptos sobre datos RDF. SWSE [HHU+11] es otro motor de búsqueda que opera directamente sobre datos RDF y consiste en un rastreo, mejora de datos, indexación y una interfaz de usuario para mostrar los resultados. Con todos estos navegadores semánticos disponibles, otros estudios se han centrado en aplicar nuevos enfoques para clasificar la información obtenida de la búsqueda como se propone en [ADA16]. Como hemos señalado anteriormente, hay un segundo grupo de aplicaciones de datos vinculados que corresponden a aplicaciones que cubren las necesidades de comunidades de usuarios específicos. Por ejemplo, en la categoría de medios de comunicación en la nube de datos vinculados, uno de los más populares es la plataforma de datos vinculados de la BBC. La BBC es una de las primeras organizaciones en utilizar datos vinculados mediante el apoyo a eventos relacionados con deportes, educación, música, etc. mediante un conjunto de ontologías como la ontología de la BBC, la ontología de los alimentos, etc. En la categoría de dominio cruzado en La Web de datos, un proyecto interesante es IBM Watson [IBM], que es una plataforma de AI capaz de comprender preguntas complejas y responderlas. La plataforma incorpora bases de conocimiento que siguen los estándares de la tecnología Linked Data. Otro ejemplo en la categoría de la red social es

la API Facebook graph, un proyecto que presenta una visión del grafo social de Facebook [WT13] donde los perfiles de los usuarios se presentan como nodos y sus relaciones como arcos. Esta información de salida formateada como JSON se convirtió a formato Turtle, que es semánticamente más rico que el formato JSON. Finalmente, Google [STH10] también utiliza formatos de datos vinculados para mostrar fragmentos enriquecidos para páginas web, lo que mejora la forma en que se muestran los resultados y tiene un impacto importante en términos de economía.

En este capítulo, hemos definido el concepto de aplicaciones de datos vinculados. También los hemos clasificado en dos grupos llamados aplicaciones de datos vinculados genéricos o de dominio específico. Se han proporcionado algunos ejemplos de estas aplicaciones de datos vinculados para cada grupo. En el siguiente capítulo, describimos cómo los Datos Vinculados en el campo de la Biodiversidad y Oceanografía han evolucionado en la última década, las aplicaciones de Datos Vinculados implementadas durante este período, y analizaremos los requerimiento que se definieron para el sistema *BiGe-Onto*.

Capítulo 3

Aplicaciones Especificas del Dominio

Índice

En este capítulo, presentamos el impacto y los usos que los datos enlazados han tenido en la Biodiversidad y Oceanografía. Luego, destacamos la importancia de las ontologías en estas áreas, centrándose en la falta de un sistema capaz de gestionar de forma integrada información de estas disciplinas. Finalmente relevamos una serie de requerimientos que el sistema *BiGe-Onto* debe cumplir para contribuir para el manejo integrado de información.

3.1. Datos Enlazados en Oceanografía y Biodiversidad

El campo de las ciencias de la vida ha entrado en una nueva era de Big Data con los avances en ciencias y tecnología. Este hecho ha incrementado el interés de la comunidad científica en las tecnologías de Datos Vinculados. Estas tecnologías permiten a los usuarios aplicar estos estándares a un sin fin de datos, como biológicos, de Biodiversidad y Oceanografía, para luego integrarlos y publicarlos como Datos Enlazados.

3.1.1. Datos Enlazados en Oceanografía

Existen varios proyectos que hacen uso de los Datos Enlazados en el campo de la Oceanografía. El primero de ellos denominado *Rolling Deck to Repository (R2R)* [ACS+13] financiado por NSF (U.S. National Science Foundation) para proporcionar la administración de datos de sensores ambientales rutinariamente recopilados por la flota de investigación académica de Estados Unidos trabajando en estrecha colaboración con el Sistema de Laboratorio Oceanográfico Universitario Nacional (UNOLS) y los Centros Nacionales de Datos NOAA. R2R mantiene un catálogo de embarcaciones, sistemas de instrumentos, expediciones, conjuntos de datos, investigadores, organizaciones, adjudicaciones de fondos, informes de cruceros y vías de navegación. Cada crucero oceanográfico financiado por NSF crea registros en R2R para su posterior publicación. Como tal, R2R garantiza la preservación y el acceso a los recursos de datos oceanográficos nacionales de EE.UU. Y proporciona una puerta de entrada central a través de la cual los datos de las expediciones oceanográficas se catalogan de forma rutinaria y se transmiten de manera segura a los archivos nacionales a largo plazo, incluido el Centro Nacional de Datos Geofísicos (NGDC) y el Centro Nacional de Datos Oceanográficos (NODC). R2R proporciona así la documentación de datos esenciales para cada expedición y herramientas para mejorar la documentación de la amplia gama de actividades de adquisición de datos a bordo

típicas de las expediciones modernas. Además, proporciona una evaluación de la calidad posterior a los cruceros de los datos y la retroalimentación a los operadores. La estructura del grafo RDF subyacente en el conjunto de datos vinculado a R2R utiliza un conjunto de patrones de diseño de ontología interrelacionados que se describen en [NF15]. En la Figura 3.1 se puede apreciar una vista conceptual del esquema.

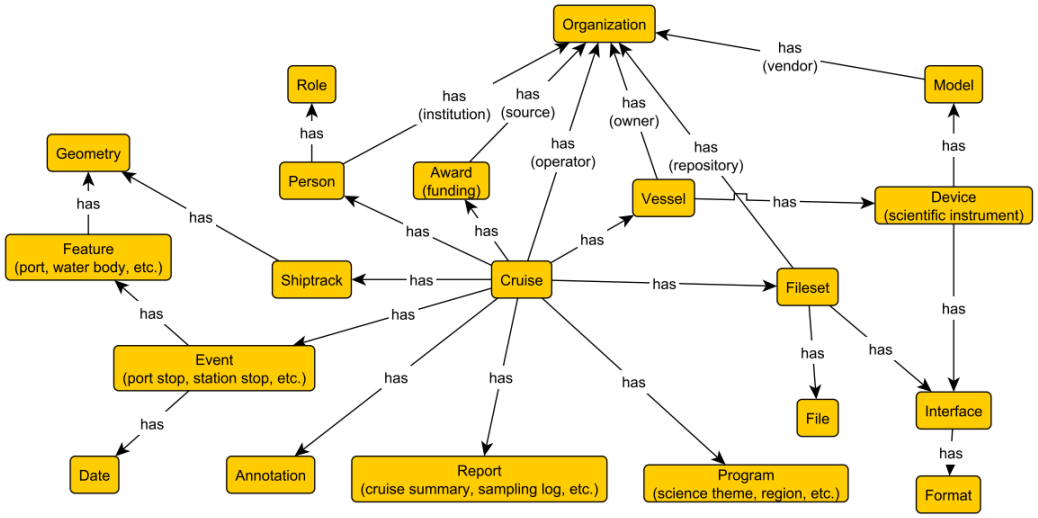

Figura 3.1: Diagrama conceptual del esquema de R2R.

El conjunto de datos vinculados R2R actualmente consta de más de 780 mil tripletas, a las que se puede acceder a través de su SPARQL endpoint `http://data.rvdata.us/snorql/`. Los metadatos legibles por máquina están disponibles en `http://data.rvdata.us/.well-known/void`. También se proporciona una URL para los navegadores web semánticos `http://data.rvdata.us/all`. El SPARQL endpoint se alimenta de la base de datos interna R2R y, por lo tanto, es actual. La descarga masiva es posible a través de `http://www.rvdata.us/outgoing/lod/`

Otro caso de éxito a nivel internacional es la Oficina de Gestión de Datos Oceanográficos Biológicos y Químicos (*Biological and Chemical Oceanography Data Management Office*, BCO-DMO) [BCO] que fue creada para servir a los investigadores financiados por los Programas de Oceanografía

Biológica, Oceanografía Química y Organismos Antárticos y Ecosistemas de la Fundación Nacional de Ciencias como una instalación donde se elaboran datos e información biogeoquímica y ecológica marina en el curso de la investigación científica que se puede diseminar, proteger y almacenar fácilmente en plazos cortos e intermedios. La Oficina de Administración de Datos también proporciona a los investigadores y a otras personas las herramientas y sistemas necesarios para trabajar con datos biogeoquímicos y ecológicos marinos de fuentes heterogéneas con una mayor eficacia. El sistema de datos BCO-DMO puede acomodar muchos tipos diferentes de datos, incluyendo mediciones y resultados biológicos, químicos y físicos. El sistema proporciona acceso a los datos (números, imágenes y/o documentos) de manera consistente, con suficientes metadatos, de modo que otros puedan hacer un uso completo de estos datos para sus propios fines. La existencia de suficientes metadatos permite el descubrimiento y la reutilización precisa de los datos por parte de los investigadores iniciales que recopilan y procesan los datos. BCO-DMO utiliza una ontología diseñada manualmente para la organización de datos, descripta en [CGS+13]. El diagrama conceptual se puede ver en la Figura 3.2.

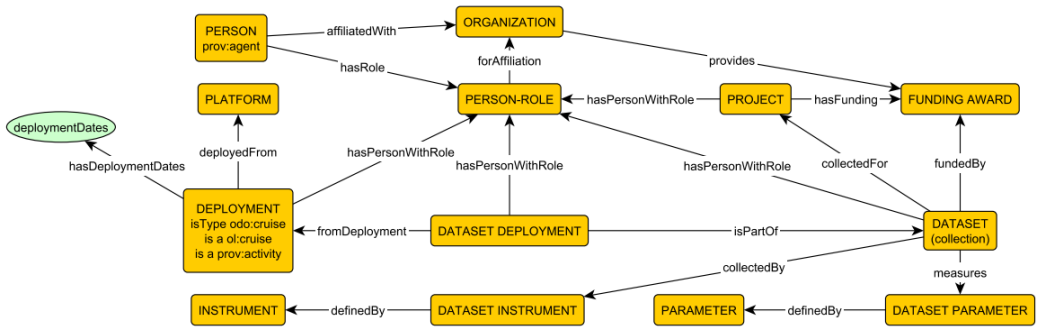

Figura 3.2: Diagrama conceptual del esquema BCO-DMO.

El conjunto de Datos Enlazados BCO-DMO tiene metadatos legibles por máquina accesibles a través de `http://www.bco-dmo.org/.well-known/void`. Todo el conjunto de datos consta actualmente de más de 8 millones de tripletas. Se puede acceder a las mismas a través de un SPARQL endpoint y se proporciona una interfaz visual para explorar el conjunto de datos en

http://www.bco-dmo.org/linked-open-data. El conjunto de datos se alimenta de la base de datos interna de BCO-DMO y, por lo tanto, se encuentra actualizada.

Finalmente, un proyecto muy interesante que hace uso de los Datos Enlazados, es el Knowledge Graph de GeoLink [CKA+18], proyecto financionado por la iniciativa EarthCube[1], la cual ha aprovechado los principios de Datos Enlazados para crear un conjunto de datos que permite a los usuarios consultar y razonar en algunos de los repositorios de metadatos de geociencia más destacados de los Estados Unidos. El conjunto de datos de GeoLink incluye información tan diversa como paradas en puertos realizadas por cruceros oceanográficos, metadatos de muestras físicas, financiamiento de proyectos de investigación y personal, y autoría de informes técnicos. Los datos se han publicado de acuerdo con las mejores prácticas para los Datos Enlazados y están disponibles públicamente a través de un punto final SPARQL que en la actualidad contiene más de 45 millones de tripletas RDF junto con una colección de ontologías y herramientas de visualización geográfica.

3.1.2. Datos Enlazados en Biodiversidad

Uno de los proyectos mas relevantes que podemos mencionar es el grafo de conocimiento denominado Ozymandias [Pag19] el cual utiliza identificadores compartidos para vincular información abierta disponible sobre taxones, revistas, publicaciones y personas, este grafo de conocimiento de la biodiversidad revela las conexiones entre los investigadores, los resultados de la investigación y los datos que respalda su trabajo. El prototipo inicial de Ozymandias combina una clasificación de animales extraídos del la base de datos Atlas of Living Australia[2] con datos sobre nombres taxonómicos y publicaciones del directorio de fauna australiana. Identificadores comunes como DOI (Identificadores de Objetos Digitales) para artículos, LSID (Identificadores de Ciencias de la Vida) para nombres taxonómicos y ORCID (Open Re-

[1] https://www.earthcube.org/group/geolink
[2] https://www.ala.org.au/

searcher and Contributor ID) para identificar personas. Ozymandias permite a los usuarios ocasionales, estudiantes e investigadores explorar millones de relaciones y utilizar cualquier parte como punto de entrada para la investigación, centrándose en el conocimiento de una especie en particular, el rango de actividades de un investigador o el resultado de investigación asociado con una institución en particular.

Los grafos de conocimiento como Ozymandias tienen numerosas aplicaciones prácticas, como informar la recopilación de datos y las políticas de gestión, por ejemplo, al descubrir las brechas en la digitalización de la literatura o la representación desigual del contenido de diferentes instituciones. También proporcionan una herramienta crítica para involucrar a los investigadores taxonómicos en la curación de datos, debido a que si un investigador tiene un ORCID, podemos descubrir su lista de publicaciones e identificar los taxones en los que trabajan sin tener que pedir demostraciones de su experiencia.

3.2. Ontologías y vocabularios para Oceanografía y Biodiversidad

En esta sección presentamos las principales ontologías y vocabularios utilizados en el dominio de Biodiversidad y Oceanografía que se tuvieron en cuenta para desarrollar *BiGe-Onto*. Siempre que fue posible, intentamos utilizar vocabularios y ontologías recomendados por W3C para garantizar la reutilización y la interoperabilidad.

3.2.1. Ontologías y vocabularios para Biodiversidad

Existen importantes aplicaciones en curso que hacen uso de las tecnologías semánticas en las ciencias biológicas, como la Ontología de Colecciones Biológicas (BCO) [WDG+14]. El objetivo del BCO es apoyar la inter-

operabilidad de los datos de biodiversidad, incluidos los datos sobre colecciones de museos, muestras ambientales/metagenómicas y estudios ecológicos. En el dominio de la ecología, podemos destacar la Ontología Ambiental (EnvO) [BMS+13] que proporciona un vocabulario controlado y estructurado diseñado para respaldar la anotación de cualquier organismo o muestra biológica con descriptores ambientales. EnvO también contiene términos para biomas, características ambientales y material ambiental.

Darwin Core (DwC) [WBG+12] es un cuerpo de estándares para la biodiversidad. Proporciona términos y vocabularios estables para compartir datos de biodiversidad. DwC es de mantenimiento de TDWG (Estándares de información sobre biodiversidad, anteriormente Grupo internacional de trabajo sobre bases de datos taxonómicos)[3]. Sus términos están organizados en nueve categorías (a menudo denominadas clases), siete que cubren amplios aspectos del dominio de la biodiversidad. Las categorías restantes cubren relaciones con otros recursos, mediciones e información genérica sobre los registros. Especialmente para el nivel récord DwC recomienda el uso de una serie de términos de Dublin Core (tipo, modificación, idioma, derechos de los titulares, derechos de acceso, citas bibliográficas, referencias). El conjunto completo de términos DwC actuales con sus descripciones está disponible en la Guía de referencia rápida `http://rs.tdwg.org/dwc/terms/`.

El esquema de acceso a datos de colecciones biológicas (ABCD, por sus siglas en inglés) [ABC19] es un estándar integral en evolución para el acceso e intercambio de datos sobre especímenes y observaciones (datos primarios de biodiversidad). El esquema ABCD trata de ser completo y altamente estructurado, apoyando datos de una amplia variedad de bases de datos. Es compatible con varios estándares de datos existentes. Existen estructuras paralelas para que (o ambos) datos atomizados y texto libre puedan ser acomodados. Las versiones 1.2 y 2.06 están actualmente en uso con las redes GBIF y BioCASe (Servicio de acceso de recolección biológica para Europa).

Si bien no es una ontología de Biodiversidad, nos permite representar

[3] `http://www.tdwg.org/`

información geoespacial para datos biológicos u oceanográficos. Estamos hablando de GeoSPARQL [PH12] estándar del Open Geospatial Consortium (OGC) que admite la representación y consulta de datos geoespaciales en la Web Semántica (SW) [BLHL+01]. Como tal, se basa en el modelo de características simples de OGC, con algunas adaptaciones para RDF. GeoSPARQL designa un vocabulario para representar datos geoespaciales en RDF, y define una extensión del lenguaje de consulta SPARQL para procesarlos, junto con una pequeña ontología para representar características[4] y geometrías[5], y una serie de predicados y funciones de consulta SPARQL. Todas estas definiciones se derivan de otras normas de OGC para que estén bien fundamentadas y documentadas. El uso de la nueva norma debería garantizar dos cosas: (1) si un proveedor de datos utiliza la ontología espacial en combinación con una ontología de su dominio, esos datos se pueden indexar y consultar correctamente en los almacenes RDF espaciales; y (2) las tiendas triples RDF compatibles deben poder procesar correctamente la mayoría de los datos RDF espaciales. Aunque existen vocabularios alternativos, como GeoRDF [Geo11] que permite representar datos simples como la latitud, la longitud y la altitud como propiedades de puntos (usando WGS84 como referencia) y GeoOWL [Geo07], que permite expresar objetos espaciales (líneas, rectángulos, polígonos), optamos por GeoSPARQL porque ofrece buenas capacidades de razonamiento espacial para comparar geometrías. W3C Time Ontology [HP06] proporciona un vocabulario para expresar hechos sobre relaciones topológicas entre instantes e intervalos, junto con información sobre la duración y sobre la información de fecha y hora. Esta ontología proporciona un vocabulario para expresar hechos sobre relaciones topológicas (ordenación) entre instantes e intervalos, junto con información sobre duraciones y sobre posición temporal, incluida información de fecha y hora. Las posiciones de tiempo y las duraciones se pueden expresar utilizando el calendario y el reloj convencionales (gregorianos), o usando otro sistema de referencia temporal como el tiempo de Unix, el tiempo geológico o diferentes calendarios.

[4]Una característica es simplemente cualquier entidad en el mundo real con alguna ubicación espacial.

[5]Una geometría es cualquier forma geométrica, como un punto, un polígono o una línea, y se utiliza como una representación de la ubicación espacial de una entidad.

3.2.2. Ontologías y vocabularios para Oceanografía

El Servidor de Vocabulario del Consejo de Investigación del Ambiente Natural (NERC) [LLC12] proporciona acceso a listas de términos estandarizados que cubren un amplio espectro de disciplinas relevantes para la comunidad oceanográfica y más amplia. El uso de conjuntos estandarizados de términos (también conocidos como vocabularios controlados) en metadatos y etiquetas de datos resuelve el problema de ambigüedades asociadas con el marcado de datos y también permite que los registros sean interpretados por las computadoras. Esto abre los conjuntos de datos a todo un mundo de posibilidades para la administración asistida por computadora, la distribución y la reutilización a largo plazo. Por ejemplo, a veces pueden ocurrir errores de nivel de datos, que son causados por diferencias que ocurren en dominios de datos debido a múltiples representaciones posibles, interpretaciones de datos similares o incluso errores de ortografía, por ejemplo. *Oxígeno, O2, Oxgen.* Para el ojo humano, la similitud es obvia, pero una computadora no podría interpretar esto como lo mismo a menos que todas las opciones posibles estuvieran codificadas en su software. Si los datos están marcados con los mismos términos, este problema se resuelve.

Geolink [KAC+14] describe un patrón de diseño ontológico (ODP) para cruceros oceanográficos utilizando OWL. Este patrón se especificó como una combinación y reutilización de los patrones existentes: trayectoria, evento y objeto de información. Consideramos que este ODP es suficientemente genérico y se adapta bien a nuestros requisitos, y por este motivo se adoptará para definir las relaciones y clases que designaremos en nuestro conjunto de datos.

Dentro del contexto nacional, el Ministerio de Ciencia, Tecnología e Innovación Productiva, desarrolló el sistema nacional de Datos Biológicos (SNDB)[6] sobre una plataforma denominada Integrated Publishing Toolkit [RDG+14] desarrollada por GBIF para datos biológicos basados en el estándar Darwin

[6] http://www.datosbiologicos.mincyt.gob.ar/

Core [WBG+12]. Para datos marinos se creó el sistema nacional de datos del mar (SNDM)[7] sobre la plataforma desarrollada por International Oceanographic Data and Information Exchange (IODE), con el fin de visualizar la información de los centros nacionales productores de datos marinos de Argentina de tipo satélites, geológicos, batimétricos, físico-químicos, geofísicos y biológicos. En la actualidad este portal presenta limitaciones, como por ejemplo: no permite visualizar de manera gráfica campañas oceanográficas u otro tipo de información de manera interactiva, convirtiéndolo en una herramienta poco atractiva para los investigadores o público en general interesado en la oceanografía. No permite interoperar con otros repositorios de datos científicos. Por otra parte existe un desinterés en contribuir con conjuntos de datos por parte de la comunidad científica, (en 2018 solo se registraron dos conjuntos de datos nuevos).

3.3. Requerimientos de *BiGe-Onto*

A través de una revisión exhaustiva de la bibliografía, se relevaron las iniciativas mas importantes de datos enlazados, ontologías y vocabularios del dominio oceanográfico y de biodiversidad. Si bien muchas de esta iniciativas funcionan para sus dominios específicos, en la actualidad no existe un sistema capaz de manejar la información de manera integral. Por ejemplo no existe un sistema capaz de hacer coincidir las ocurrencias de especies con las variables ambientales, esto es un requisito fundamental de los análisis macroecológicos [MKMB+18], en particular aquellos que consideran los impulsores ambientales de la distribución de especies, y cómo se espera que las distribuciones cambien según el clima cambios [PD03]

Estas consideraciones proporcionan una fuerte motivación para formular un sistema que tenga en cuenta las características geoespaciales que pueden brindar respuestas a preguntas como: (i) ¿Cómo podemos definir las regiones espaciales para nuestros estudios? (ii) ¿Cómo se distribuyen las especies en

[7] http://www.datosdelmar.mincyt.gob.ar/

una determinada región? (iii) Dada una georeferencia particular, ¿a qué región geográfica pertenece? (iv) ¿Cómo relacionar las ocurrencias de especies con variables ambientales dentro de una región específica?.

Si bien el principal requerimiento tiene que ver con la información geoespaciales definimos una serie de requerimientos complementarios, los cuales se detallan a continuación y se desarrollan en profundidad en los siguientes capítulos de esta investigación.

- **(R1) Soporte geoespacial**: el sistema tiene que poder responder consultas que impliquen operaciones con relaciones topológicas entre geometrías definas por el usuario.
- **(R2) Interoperabilidad con otros sistemas** es un requerimiento esencial para un manejar y compartir información o complementarla con otros sistemas.
- **(R3) Interfaz visual**: para que los expertos del dominio puedan interpretar de manera intuitiva los resultados.
- **(R4) Proveer un punto de acceso**: es vital poder compartir la información de manera que una maquina la puede interpretar y representar en diferentes formatos.
- **(R5) Metodología formal**: el modelo conceptual subyacente al sistema tiene que seguir una metodología formal de desarrollo para asegurar su validez, modularidad y consistencia.
- **(R6) Cobertura de ambos dominios**: el sistema y el modelo propuesto debe se capaz de representar información de Biodiversidad y Oceanografía de manera transparente para el usuario.

Es indudable que los Datos Enlazados generaron en los últimos años gran interés en las comunidades científicas, en este capítulo vimos aplicaciones relacionadas con Biodiversidad y Oceanografía y sus estándares mas utilizados.

Sin embargo debido a que estas aplicaciones no cubren las necesidades de un sistema que maneje de manera integrada la información de estas disciplinas, definimos una serie de requisitos a cubrir por el sistemas a desarrollar. En el siguiente capítulo, veremos la importancia de las ontologías (OWL) como una extensión de RDF en estas dos disciplinas y como podemos utilizarla para desarrollar un sistema para el manejo integrado de esta información denominado *BiGe-Onto*.

Capítulo 4

Desarrollo de BiGe-Onto

Índice

En este capítulo, se presenta la primer versión de *BiGe-Onto* [ZBF+19] disponible en linea a partir de Noviembre de 2018 en `http://crowd.fi.uncoma.edu.ar/cenpat-gilia/bigeonto/` y el conjunto de datos enlazados publicado en `https://doi.org/10.5281/zenodo.3235548`. Asimismo, aborda y justifica los requerimientos (R2) Interoperabilidad con otros sistemas, (R3) Visualización de datos, (R4) Proveer un punto de acceso y (R5) Metodología formal.

En [STH+19] se presentaron algunos de los resultados referidos al requisito (R2), que posteriormente fueron tomados como referencia para esta investigación, mientras que en [ZBF] se presentan resultados referidos al requisito (R4) y que de la misma manera fue incorporados aquí.

4.1. *BiGe-Onto* visión general

Recientemente se ha logrado un gran progreso para digitalizar los datos de Biodiversidad y Oceanografía disponibles en el mundo, pero el manejo de los datos de muchos proveedores diferentes y en todos los dominios de investigación sigue siendo un desafío. Una revisión del panorama actual de estándares de metadatos y ontologías en las ciencias de la vida, sugiere que los estándares existentes, como Darwin Core [WBG+12] son inadecuados para describir los datos de biodiversidad de forma computacionalmente útil. Como contribución para llenar este vacío, presentamos un sistema basado en ontologías, llamado *BiGe-Onto* [ZBF+19], diseñado para administrar datos de manera integrada para el dominio de Biodiversidad y Biogeografía marina.

Como fuentes de datos, utilizamos dos bases de datos ampliamente utilizadas en estas dos disciplinas. Una de las bases de datos de Biodiversidad más reconocidas en todo el mundo es el Fondo de Información de Biodiversidad Global (GBIF) [GBI19], que es una red internacional de investigación financiada por los gobiernos del mundo y orientada a proporcionar acceso abierto a datos sobre todos los tipos de vida en la Tierra. La misión de GBIF es ser el principal recurso global para la información sobre la biodiversidad y generar soluciones inteligentes para el bienestar ambiental y humano. Para lograr esta misión, GBIF alienta a una amplia variedad de proveedores de datos en todo el mundo a entregar datos a través de su red. Por otro lado, el Sistema de Información Biogeográfica del Océano (OBIS) [OBI19] es una federación internacional de organizaciones y personas que comparten una visión para hacer que los conjuntos de datos biogeográficos marinos de todo el mundo estén disponibles gratuitamente a través de Internet. Funciona como un proveedor basado en la web de información georreferenciada global sobre especies marinas.

Desde la perspectiva de las bases de datos, se pueden encontrar algunas similitudes entre OBIS y GBIF. Lo más destacable es que ambos publican sus conjuntos de datos a través del Kit de herramientas de publicación integrado

(IPT) [RDG+14], que es una herramienta de software de código abierto que se utiliza para publicar y compartir los conjuntos de datos de Biodiversidad y Biogeografía. Estos conjuntos de datos están disponibles como Darwin Core Archives (DwC-A) [RD11], que consiste en un conjunto de archivos para describir la estructura y las relaciones de los datos sin procesar junto con los archivos de metadatos que utilizan el estandard Darwin Core [WBG+12]. Aunque ambas bases de datos comparten el mismo estándar, OBIS tiene particularidades relacionadas con la naturaleza de sus datos (sistemas marinos), por ejemplo, tiene más que solo la ocurrencia de especies. Estos datos se recopilan como parte de la investigación biológica marina que incluye mediciones de las características del hábitat, como las variables físicas y químicas del entorno, y mediciones biométricas (como el tamaño corporal, conteos, abundancia y biomasa combinadas, etc.), así como detalles. con respecto a la naturaleza del muestreo o los métodos de observación, el equipo y el esfuerzo de muestreo.

Con este fin, proponemos el diseño de un sistema basado en onotologias llamado *BiGe-Onto* compuesto de (i) Arquitectura (ii) un modelo conceptual, (iii) una versión operacional de *BiGe-Onto* codificada en OWL 2, y (iv) un conjunto de datos enlazados disponible para su explotación a través de un punto final SPARQL. Además, mostraremos casos de uso que permiten a los investigadores responder preguntas que manejan información de ambos dominios.

4.2. Arquitectura

Un patrón arquitectónico importante utilizado en el desarrollo de sistemas, es la arquitectura *multi-niveles* [Sch09]. Una arquitectura de varios niveles separa la funcionalidad en varias capas, desde el almacenamiento de datos de bajo nivel hasta los componentes de interacción del usuario. Esta arquitectura se usa comúnmente para muchos tipos de aplicaciones web. Como muchas aplicaciones de datos vinculados también son aplicaciones web,

tienden a comprender este enfoque arquitectónico. Una ventaja importante de la arquitectura multi-nivel es que separa lógicamente la funcionalidad del sistema en una serie de capas y especifica la comunicación entre esas capas. Esta separación hace que sea mucho más fácil reemplazar una capa de la arquitectura o reutilizar una capa de una arquitectura existente en una nueva aplicación. La arquitectura multi-niveles más utilizada es la arquitectura de tres niveles [Eck95]. Debido a su simplicidad y probada confiabilidad, decidimos basar nuestra arquitectura en este modelo. La Figura 4.1 ilustra el diseño de la arquitectura *BiGe-Onto* y las siguientes secciones describen cada una de las capas.

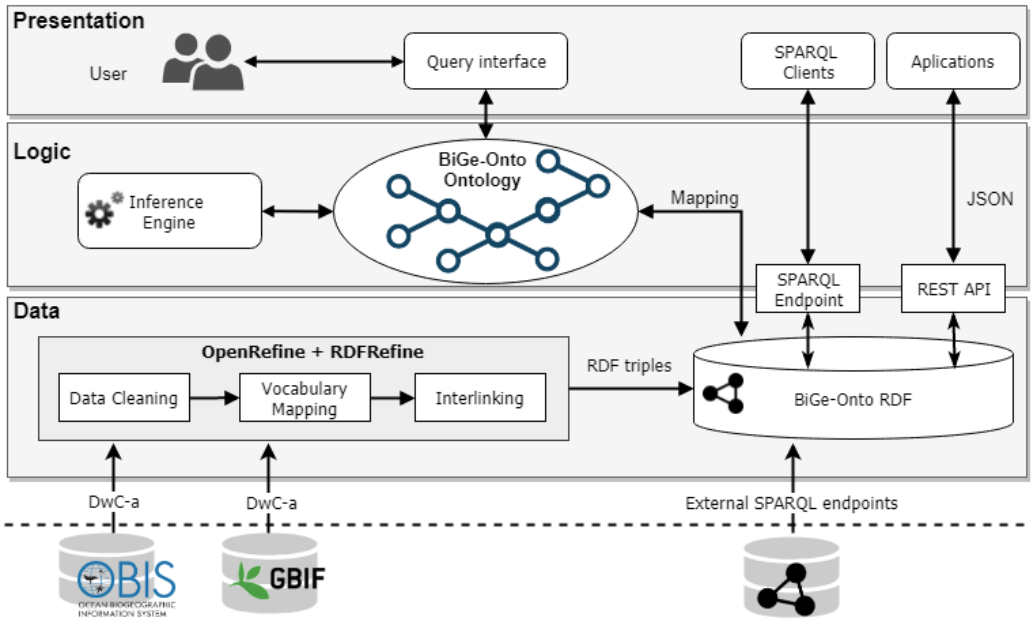

Figura 4.1: Arquitectura de tres capas utilizada por *BiGe-Onto*.

Capa 1 (Datos de entrada): la capa de datos almacena los datos subyacentes independientemente de la lógica de negocios. En este caso, los conjuntos de datos que pertenecen a OBIS y GBIF en DwC-A se transforman a RDF y se exportan posteriormente en formato Turtle. Después de eso, se importan a GraphDB (RDF triplestore) que admite diferentes serializaciones RDF. GraphDB permite a los usuarios explorar la jerarquía de las clases RDF, donde se puede examinar cada clase para examinar sus instancias. De

manera similar, las relaciones entre estas clases también se pueden explorar, dando una visión general de cuántos enlaces existen entre las dos instancias de clases. Cada enlace es una declaración RDF donde el sujeto y objeto son instancias de clase y su predicado es el enlace entre estos. Por último, los usuarios también pueden explorar recursos que proporcionen URI que representen a cualquiera de los sujetos, predicados u objetos de una tripleta. El modelo basado en tripletas de *BiGe-Onto* está diseñado para persistir, ya que puede exportarse completamente en RDF e importarse a otra solución compatible con RDF. Es importante tener en cuenta que también podemos importar datos a GraphDB desde puntos finales SPARQL que permitan consultas federadas.

De esta manera cumplimos con el requisito (R4) Proveer un punto de acceso, para que otras aplicaciones puedan consumir nuestros datos.

Capa 2 (Lógica): una vez que los datos integrados están disponibles en GraphDB, pueden ser utilizados y accedidos por las capas lógica y de presentación. Parte de la lógica puede implementarse en la capa de datos al razonar sobre el repositorio RDF, aunque las posibilidades de razonamiento son limitadas porque no permite la definición de restricciones de propiedad de dominio y rango o definición de relaciones jerárquicas. Por lo tanto, se requiere un mayor nivel de expresividad para lograr un razonamiento adecuado. En esta capa, la ontología permite la identificación inequívoca de entidades y la afirmación de las relaciones con nombre aplicables que conectan estas entidades. Específicamente, *BiGe-Onto* cumple los siguientes roles: (i) Explicación de contenido: la ontología permite la interpretación precisa de datos de múltiples fuentes a través de la definición explícita de términos y relaciones, (ii) Modelo de consulta: la consulta es formulada utilizando la ontología como un esquema de consulta global, y (iii) Verificación: la ontología verifica las asignaciones utilizadas para integrar datos de múltiples fuentes.

Capa 3 (Presentación): una de las características proporcionadas por GraphDB, es una interfaz de tipo asistente que dirige a los usuarios en la creación de varias visualizaciones de datos RDF con diferentes puntos de

partida. El usuario puede configurar la pantalla gráfica predeterminada con la expresividad completa del lenguaje SPARQL para controlar qué datos gráficos desea mostrar. GraphDB permite resolver muchos de los problemas complejos que surgen al tratar con datos de ciencias de la vida. Esto permite controlar el punto de inicio de la visualización y crear más de una visualización en la misma información. Con esta herramienta, la exploración y el análisis de datos, y el descubrimiento de conocimientos se vuelven más fáciles y rápidos. Las herramientas GraphDB se pueden usar para inferir relaciones que no están establecidas explícitamente, para obtener una imagen completa de los datos y mejorar el conocimiento sobre los enlaces en los conjuntos de datos. La Figura 4.2 muestra un ejemplo de una visualización GraphDB, que le permite al usuario ver una instancia de la clase *Taxon* que representa una especie en particular.

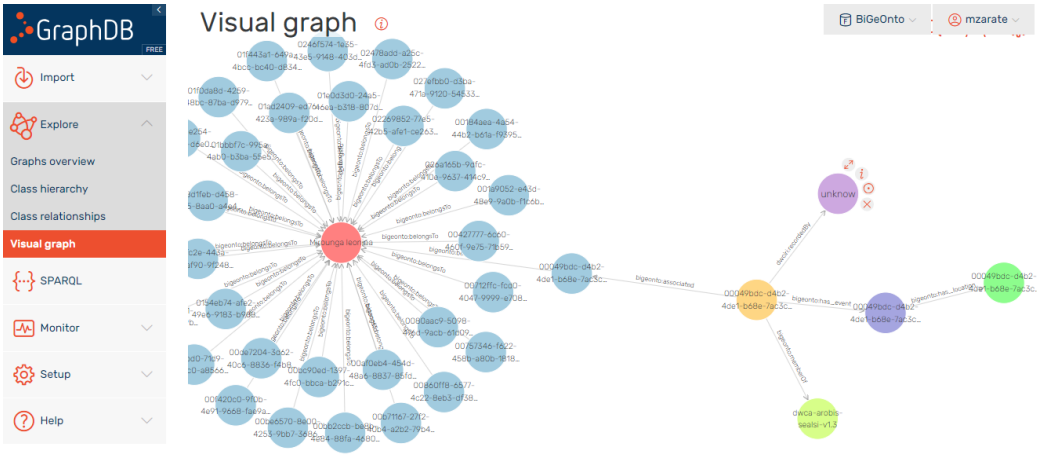

Figura 4.2: Visualización en forma de grafo utilizando GraphDB. El circulo rojo representa el taxon (*Mirounga Leonina*) y las relaciones existentes con la ubicación, las personas, el conjunto de datos de origen, etc.

De esta manera cumplimos parcialmente con el requisito (R3) Visualización de datos. Consideramos que no es un requisito cumplido, debido a que es necesario tener conocimientos de RDF, SPARQL y algunos conceptos relacionados con la Web Semántica para interpretar la información, por lo que este es uno de los requisitos que necesitará de mayor desarrollo en el futuro para lograr una interfaz visual amigable para los investigadores de otras

áreas.

4.3. Modelo conceptual

Siguiendo las prácticas de ingeniería de software, presentamos un análisis basados en una visión general (ontología como modelo conceptual OntoUML [Gui05]) y posteriormente adoptamos la ontología a un lenguaje de implementación como lo es OWL 2. OntoUML se basa en la Ontología Fundamental Unificada (Unified Foundational Ontology, UFO) [Gui05, GW10], que es un lenguaje de modelado ontológicamente bien fundado, mientras que Menthor [MSG+16] es un editor visual ligero para este lenguaje, permitiendo a los usuarios lidiar con la complejidad de sus principios ontológicos subyacentes. Al mismo tiempo, refuerza estos principios en los modelos producidos al proporcionar un mecanismo para la verificación formal automática de restricciones [BG09].

Para construir el modelo conceptual, consideramos un enfoque sistemático, llamado SABiO [dAF14], que se centra en el desarrollo de ontologías de referencia de dominios específicos. Las ontologías de referencia de dominio proporcionan modelos conceptuales independientes de la solución para hacer la mejor descripción posible del dominio. Una vez que los usuarios están de acuerdo con estas ontologías de referencia, se pueden implementar ontologías legibles por máquina que garanticen las propiedades computacionales deseables. Las actividades principales de SABiO son las siguientes: (i) **Identificación del propósito de ontología y elicitación de requisitos**: el resultado principal de esta fase es un conjunto de preguntas de competencia (Competency Questions, CQ), que ayuda a capturar la conceptualización del dominio; (ii) **Captura y formalización de ontologías**: esta fase depende del conjunto de CQ identificados anteriormente, generando un diagrama OntoUML, que se basa en UFO, y un conjunto de axiomas formales (si los hay); (iii) **Diseño de ontología**: esta fase tiene como objetivo transformar la ontología de referencia en una ontología operacional (legible por máquina)

expresada en un lenguaje de ontología estándar como OWL 2; (iv) **Implementación de la ontología**: la única actividad de esta fase consiste en implementar la ontología en el lenguaje operativo elegido; y (v) **Pruebas de ontología**: la fase de prueba de SABiO es impulsada por la preguntas de competencia, es decir, validación y verificación de la ontología operacional en un conjunto de pruebas derivadas de estos preguntas (Ver capítulo 5). Las pruebas se implementan como consultas en la ontología instanciada. Además, las ontologías operativas pueden verificarse para ver si son satisfactorias conceptualmente invocando sistemas de razonamiento disponibles.

El desarrollo de *BiGe-Onto* sigue las mejores prácticas para la construcción de ontologías, como la reutilización de los conceptos de ontologías mostradas en la Sección 3.2 y directrices conocidas de SABiO basadas en los conceptos principales de OntoUML. Las ontologías que componen *BiGe-Onto* se han implementado en OWL 2 siguiendo las mejores prácticas para publicar datos enlazados [JHA+14]. Por lo tanto, las ontologías existentes que coinciden con nuestros requisitos fueron reutilizadas. Estas ontologías que coinciden con los requisitos de representación de conocimiento que son las siguientes:

- Para representar conceptos taxonómicos como especies, se utilizó el estándar DwC [WBG+12] y los conceptos de BCO [WDG+14].
- Para representar los metadatos sobre los conjuntos de datos RDF, se utilizó el Vocabulario VoID recomendado por W3C [ACHZ11].
- Para representar información geográfica y datos geoespaciales. La ontología GeoSPARQL [PH12] recomendada por OGC, esta nos proporciona un vocabulario para expresar información geoespacial cubriendo este requisito.
- Para representar tiempo, instantes e intervalos utilizamos la ontología recomendada por W3C Time Ontology [HP06]
- Para representar la información oceanográfica NERC [LLC12] que pro-

porciona acceso a listas de términos estandarizados que cubren un amplio espectro de disciplinas relevantes para la comunidad oceanográfica.

- El vocabulario FOAF [BM14] para representar personas, grupos, organizaciones, etc.
- EnvO [BPL+16] para describir entidades ambientales de forma precisa.

Las clases principales se muestran en la Figura 4.3, en su mayoría son reutilizadas del estándar Darwin Core, ellas son: *dwc:Occurrence*, *dwc:Event*, *dwc:Taxon* y *dwc:Organism*. *BiGe-Onto* también reutiliza *foaf:Person*, *void:Dataset* y *dcterms:Location*. Además, las siguientes clases específicas, que se reutilizan desde la ontología de EnvO, proporcionan una forma más precisa de definir un ambiente, en nuestro caso *Bige-onto:Environment* se define mediante tres clases: *biome (envo:ENVO_00000428)*: modela un ecosistema para las comunidades ecológicas residentes, *material ambiental (envo:ENVO_00010483)*: describe partes de materiales ambientales que forman el medio o parte del medio de un sistema ambiental, y *característica ambiental (envo:ENVO_00002297)*: para entidades materiales que determinan un sistema ambiental.

Con referencia a las dimensiones espaciales, utilizamos GeoSPARQL para crear sentencias que describen las relaciones topológicas (contiene, se superpone) entre una región y otras entidades con una naturaleza espacial. GeoSPARQL permite expresar tales relaciones entre dos recursos (dos geometrías o dos características) utilizando propiedades topológicas (propiedades directas) o funciones topológicas (propiedades computadas). En nuestro modelo, las clases *bige-onto:Region* y *dc:Location* se especializan en *geo:Feature*: una región es un polígono cerrado (una geometría) que representa el área geográfica. Gracias a esta especialización, podemos vincular los registros de metadatos con cualquier otra información con un componente espacial y definido como *geo:Feature*. Otras entidades en esta ontología son para describir un área delimitada tanto por la latitud como por la longitud, y por lo tanto, están completamente caracterizadas desde el punto de vista ambiental.

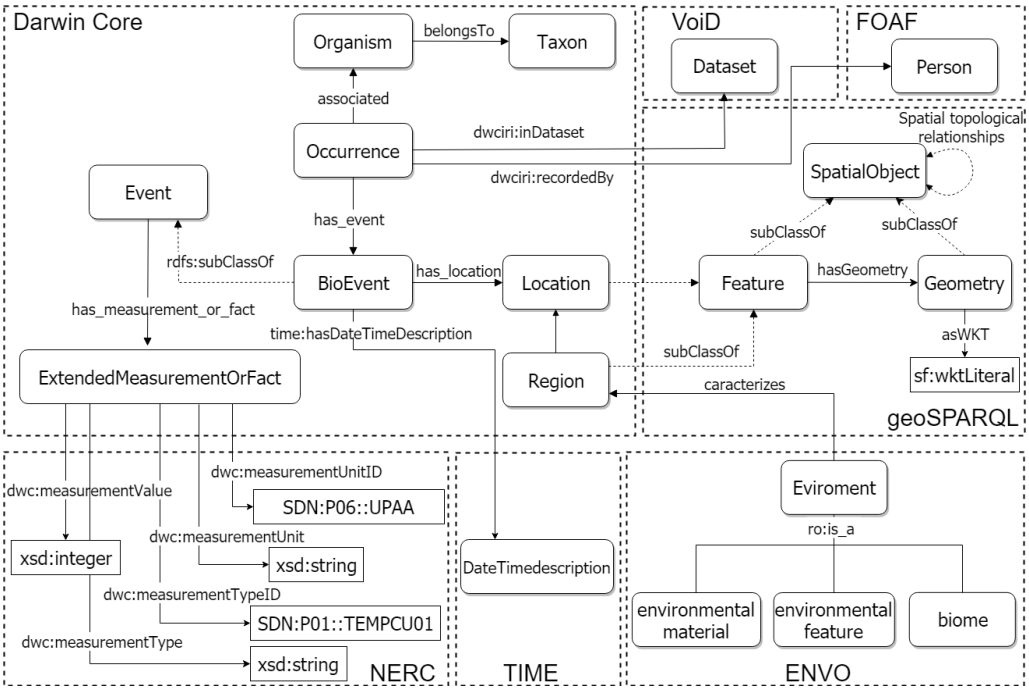

Figura 4.3: Modelo conceptual de *BiGe-Onto* .

bige-onot:BioEvent utilizada para eventos especializados asociados a Biodiversidad y Biogeografía. *BiGe-Onto* también define y reutiliza una serie de propiedades, que no se muestran en la Figura 4.3.

Dado que *BiGe-Onto* describe principalmente ocurrencias, que dependen de la existencia de otros conceptos, algunas de las siguientes propiedades más importantes se definen vinculando dichas ocurrencias con otras para proporcionar una semántica más precisa:

- *bige-onto:associated*: cada ocurrencia se describe en función de la existencia de un organismo en un lugar determinado en un momento determinado. Dicho organismo está relacionado con el taxón por medio de la relacion *bige-onto:belongsTo*.

- *bige-onto:has_event*: las ocurrencias se producen durante un evento de muestreo en una ubicación dada (*bige-onto:has_location*), que también esta caracterizada por (*bige-onto:caracterizes*) por un entorno específi-

co. Las relaciones entre *bige-onto:Environment* y las clases de EnvO están controladas principalmente por la Ontología de Relaciones (RO)[1].

- *dwciri:registeredBy*: esta propiedad proporciona información sobre personas, grupos u organizaciones que han registrado la ocurrencia. También se reutiliza desde el espacio de nombres DwC IRI y habilita rangos no literales, a diferecia de su análogo con *dwc:RecordBy*.

- *dwciri:inDataset*: esta propiedad de objeto se utiliza para vincular una ocurrencia al conjunto de datos que lo contiene.

Una de las clases más importantes de *BiGe-Onto* es *ExtendedMeasurementOrFact*, que permite modelar mediciones o hechos relacionados con una ocurrencia biológica, mediciones ambientales o información relacionada a los métodos de muestreo. Esta clase reutiliza las propiedades DwC *dwc:measurementUnitID*, *dwc:measurementTypeID* como enlaces a identificadores externos para referirse a un vocabulario para los campos de tipo de medida y unidad de medida. La dimensión temporal de un Evento identifica el momento en el que se capturó la ocurrencia, esto puede ser un instante o un período de tiempo entre fechas. Para capturar esta información reutilizamos la clase *time:DateTimeDescription* y la propiedad objeto *time:hasDateTimeDescription*. De esta manera, *BiGe-Onto* está diseñado para ser utilizado como un modelo conceptual central para ser reutilizado por diferentes grupos de investigación para diversos fines, por ejemplo, en dominios relacionados con la biología o dominios que necesitan manejar información geoespacial.

De esta forma hemos cumplido con los requisitos (R1) Soporte geoespacial y (R5) Metodología formal. La metodología formal utilizada permite que nuestro modelo sea coherente y valido para representar información de Biodiversidad y Biogeografía marina. En el capítulo 5 explicaremos en detalle como fue cubierto (R5) mediante casos de estudio.

[1] https://github.com/oborel/obo-relations

4.4. Versión operacional en OWL 2

Existe una compensación entre la expresividad y la capacidad de cálculo computacional de ontologías de referencia y operacionales. Las ontologías se utilizan como un modelo para admitir la interoperabilidad semántica, pero también como un dispositivo que debe admitir de manera eficiente el razonamiento basado en la lógica. Consideramos esta dualidad y, como también se define en SABiO, exploramos y detallamos una versión operacional OWL 2 de *BiGe-Onto* para ser instanciada con datos reales de los conjuntos de datos de Biodiversidad y Biogeografía marina. Para implementar la versión operacional de *BiGe-Onto*, seguimos las reglas de transformación establecidas por [BdSS+17] para razonar sobre características más expresivas de OWL. La jerarquía de clases y propiedades de la ontología OWL constan de las definidas previamente. También codificamos asociaciones OntoUML que relacionan las clases *BiGe-Onto* como propiedades objeto de OWL y tipos de datos OntoUML como propiedades de datos de OWL para enlazar literales. Siguiendo las reglas de transformación del metamodelo de OntoUML a OWL, cada clase de OntoUML se modela como una clase de OWL junto con generalizaciones como la definida entre *Location* y *Region*.

A diferencia de OntoUML, que no tiene asociaciones dirigidas, las propiedades OWL son relaciones binarias dirigidas. Por lo tanto, definimos una propiedad objeto para cada asociación OntoUML junto con roles inversos de OWL 2. En este contexto, las clases de origen y destino se consideran en la transformación para crear el dominio y el rango de una propiedad objeto en OWL. Por ejemplo, la relación *bige-onto:hasEvent* se define entre *dwc:Occurrence* (dominio) y *bige-onto:BioEvent* (rango). Con referencia a las cardinalidades, se transforman siguiendo los mismos criterios de [BdSS+17], aunque interpretamos la cardinalidad de OntoUML (*) como *0...N*. De manera similar, los tipos de datos se asignan a las propiedades de los datos de OWL al definir su dominio y rango respectivos. Por lo tanto, el tipo de datos *dwc:genus* es una propiedad de datos OWL con dominio *dwc:Taxon* y rango *xsd:String* (Ver Tabla 4.1). Aquí, las cardina-

lidades no se tienen en cuenta. De esta manera, *BiGe-Onto* se enfoca en el vocabulario científico y técnico requerido para lograr una comprensión semántica de los términos más utilizados en Biogeografía marina y biodiversidad. Finalmente, siguiendo las mejores prácticas para publicar ontologías, nuestra ontología está disponible a través de la URL permanente de W3ID https://w3id.org/cenpat-gilia/bigeonto/.

Cuadro 4.1: OntoUML DataTypes

Axiomas DL	**Expresión en OWL**

1. $\exists dwc : genus.\top \sqsubseteq dwc : Taxon$	`<DataPropertyDomain>`
	`<DataProperty abbreviatedIRI="dwc:genus"/>`
	`<Class abbreviatedIRI="dwc:Taxon"/>`
	`</DataPropertyDomain>`

2. $\exists dwc : genus^{-}.\top \sqsubseteq xsd : String$	`<DataPropertyRange>`
	`<DataProperty abbreviatedIRI="dwc:genus"/>`
	`<Datatype abbreviatedIRI="xsd:string"/>`
	`</DataPropertyRange>`

4.5. Conjunto de Datos Enlazados

Aunque DwC está definido en RDF, no hay herramientas proporcionadas oficialmente por TDWG para convertir archivos en formato DwC-A a RDF. Es por eso que decidimos usar OpenRefine [VDW13] para convertir los conjuntos de datos que se utilizarán para realizar las pruebas y validar nuestro modelo. OpenRefine es un paquete de software de código abierto disponible de forma gratuita para manipular conjuntos de datos, proporcionando un conjunto de herramientas fácil de usar para ayudar a extraer, revisar, transformar y exportar conjuntos de datos. Los datos se convierten a RDF utilizando RDFRefine[2], que es una extensión de OpenRefine. RDFRefine permite mapear los campos de nuestros datos a la estructura RDF requerida definiendo lo que se conoce como *Esqueleto RDF* donde se especifica cómo

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

se generarán las tripletas RDF. A diferencia de las representaciones de metadatos de valores clave, una representación de datos vinculados otorga un significado estructurado y explícito a los valores de metadatos, lo que permite que los conjuntos de datos, artículos, informes y otros recursos relacionados con la investigación se conecten mediante enlaces significativos. Sobre la base de los conjuntos de datos de GBIF y OBIS, creamos el conjunto de datos de *BiGe-Onto RDF* que contiene instancias de las diferentes clases definidas y la relación entre ellas. Por lo tanto, asignamos estos datos a la ontología siguiendo las pautas establecidas en [LKN+13].

El proceso de conversión de los archivos en formato DwC-A se describe brevemente a continuación, se puede encontrar una descripción más detallada en [BCH+07]. La primera parte del proceso implica la extracción de datos, la limpieza y la reconciliación. Los datos de los archivos DwC-A se extraen manualmente del repositorio de GBIF, en particular trabajamos con el archivo (*ocurrence.txt*, luego se procesan utilizando OpenRefine. Aquí, las ocurrencias se limpian y se convierten en tipos de datos estandarizados, como fechas, valores numéricos, etc., y se eliminan las columnas vacías si es que existen. OpenRefine también permite agregar servicios de reconciliación basados en puntos finales SPARQL, estos devuelven posibles recursos que existen en conjuntos de datos externos y que pueden coincidir con los campos de los conjuntos de datos locales. En nuestro proceso, usamos el punto final de DBpedia `https://dbpedia.org/sparql` para reconciliar la columna *País* con el recurso *dbo:country* en DBpedia, el enlace entre los recursos se realiza a través de la propiedad *owl:sameAs*. Después de esto, si la reconciliación se realizó correctamente, creamos una nueva columna para el URI correspondiente del recurso en DBPedia. En particular, agregamos la columna denominada *dbpedia:CountryURI* que nos proporcionara información adicional cuando realicemos una consulta federada por ejemplo.

Otro servicio utilizado se basa en una base de datos taxonómica llamada Enciclopedia de la vida *Encyclopedia of Life (EOL)* para reconciliar los nombres aceptados en la base de datos de EOL. Específicamente, la reconciliación se aplica al termino *dwc:scientificName*, de modo que creamos una

nueva columna llamada *EOL_page* para enlazarla a la página EOL que describe la especie, este caso la pagina no proporciona información en RDF u otro formato interpretable por maquinas, es solamente para uso informativo. Desafortunadamente, todo este proceso requiere mucho tiempo debido a que no todos los valores se combinan automáticamente y, por lo tanto, las sugerencias ambiguas deben corregirse manualmente.

Después de la limpieza y la reconciliación, la segunda parte del proceso incluye la alineación del esquema RDF y la definición de las URIs. Los datos se convierten a RDF usando RDFRefine, como mencionamos anteriormente, el esqueleto RDF especifica el sujeto, el predicado y el objeto de las tripletas que se generan. El siguiente paso en el proceso es configurar los prefijos. Dado que los conjuntos de datos incluyen localidades, ubicaciones e institutos de investigación, configuramos prefijos para vocabularios conocidos como la ontología GeoSPARQL, FOAF, Dublin Core, etc. Para generar de manera automática las URIs para cada recurso, utilizamos el lenguaje GREL (eneral Refine Expression Language) también proporcionado por OpenRefine. La estructura general de los URI es

```
http://[base_uri]/[clase DwC]/[valor]
```

donde `[base_uri]` igual a `http://www.cenpat-conicet.gob.ar/resource/`, `[clase DwC]` es la clase DwC correspondiente, y `[valor]` es el valor obtenido de la celdas en el archivo de ocurrencias. También es importante tener en cuenta que las URIs generadas son instancias de las clases definidas en el estándar DwC. En particular, dwc es una abreviatura para el espacio de nombre real `http://rs.tdwg.org/dwc/terms/`. El mapeo completo realizado junto con las columnas del archivo de texto de ocurrencia utilizado para generar los URI principales se puede consultar en el repositorio GitHub [3].

El Cuadro 4.2 resume las principales características del conjunto de datos enlazados *BiGe-Onto*.

[3] https://github.com/cenpat-gilia/BiGe-Onto/blob/master/scripts/mapping.json

Cuadro 4.2: Principales características del conjunto de datos *BiGe-Onto*.

URL del Repositorio	`http://web.cenpat-conicet.gob.ar:7200/login`
Credenciales	(user: `bigeonto` password: `bigeonto`)
SPARQL endpoint	`http://web.cenpat-conicet.gob.ar:7200/sparql`
Clases	`http://web.cenpat-conicet.gob.ar:7200/hierarchy`
Conjunto de datos	`https://doi.org/10.5281/zenodo.3235547`
Nro.Vocabularios	21
Nro. Clases	9
Nro. Propiedadess	50
Nro. triplas	4.3M

La publicación como datos abiertos enlazados permite cumplir con el requisito (R2) Interoperabilidad con otros sistemas, ya que proporciona una perspectiva para un tipo diferente de interoperabilidad, basado en los principios de la arquitectura web. La interoperabilidad de datos vinculados está diseñada para admitir modelos de descripción heterogéneos, lo cual es necesario para manejar de diferentes dominios.

En este Capítulo presentamos el desarrollo de un sistema basado en ontología denominado *BiGe-Onto* para administrar información de los dominios de Biodiversidad y Biogeografía marina, utilizando estándares como Darwin Core y GeoSPARQL. Este sistema está compuesto por (i) la arquitectura; (ii) un modelo conceptual denominado especificado en OntoUML; (iii) una versión operacional codificada en OWL 2; y (iv) un conjunto de datos integrado para su explotación a través de un punto final SPARQL.

El desarrollo y formalización de *BiGe-Onto*, justifica los requerimientos iniciales para la implementación de un sistema capaz de administrar información de manera integrada. En particular (R2) Interoperabilidad con otros sistemas, (R4) Proveer un punto de acceso y (R5) Metodología formal se cumplen completamente, mientras que requisito (R3) Interfaz visual se cumple parcialmente.

En el siguiente capítulo veremos como se evaluó *BiGe-Onto* mediante casos de estudio que permiten validar su utilidad y cumplir con el requisito (R1) Soporte geoespacial, y mediante una evaluación cuantitativa del mode-

lo conceptual propuesto. Finalmente presentamos un análisis general de los sistemas y ontologías relevadas en la literatura y que fueron evaluadas para esta investigación, junto con una descripción de cada una considerando los requerimientos propuestos para *BiGe-Onto*

Capítulo 5

Evaluación de *BiGe-Onto*

Índice

En este capitulo, expondremos el conjunto de pruebas realizadas para validar el modelo conceptual y la validación por parte de los expertos del dominio utilizando casos de uso, donde se demuestra como *BiGe-Onto* es efectivo para en el manejo integrado de información.

5.1. Validación del modelo conceptual

Como enfoque para evaluar el modelo conceptual *BiGe-Onto*, tuvimos en cuenta dos aspectos: (i) validación basada datos [BGM05], para representar situaciones reales, para validar *BiGe-Onto*, creamos ejemplos y relaciones basadas en los datos de GBIF y OBIS. (ii) pruebas ontológicas [VG06], para responder a las preguntas de competencia formuladas por expertos en el dominio. A continuación, cada uno de estos enfoques se describe en detalle.

Enfoque basado en datos para la evaluación de ontología: Para verificar el modelo propuesto, creamos instancias de diferentes conjuntos de datos extraídos de OBIS y GBIF, una instancia de sus conceptos y relaciones. El detalle completo de la instanciación se describió en la Sección 4.5. El Cuadro 5.1 muestra un ejemplo del registro contenido dentro del archivo DwC-A que pertenece al conjunto de datos titulado *Locations of seals in Patagonian Large Marine Ecosystem (OBIS South America, AR-OBIS, Sub-node)* disponible en siguiente link[1].

Cuadro 5.1: Ejemplo de instanciacion de un registo de un archivo DwC extraido desde OBIS.

Concepto	**Instancia**
Occurrence	8353bbe9-573f-4dbd-9099-c2f34c28c781
Taxon	Otaria flavescens (Shaw, 1800)
Organism	8353bbe9-573f-4dbd-9099-c2f34c28c781
Person	Lewis Mirtha
Organization	CENPAT-CONICET
BioEvent	8353bbe9-573f-4dbd-9099-c2f34c28c781
Dataset	dwca-arobis-sealsi-v1.3 Locations of seals in Patagonian Large Marine Ecosystem (OBIS South America, AR-OBIS, Sub-node)
Location	South Atlantic Ocean

Como puede verse en el Cuadro 5.1, generamos un Identificador UniversalÚnico (UUID) para cada una de las Ocurrencias usando OpenRefine. En caso de que no exista UUID persistente, se puede construir una clave combinando diferentes identificadores del registro, haciendo que el identificador del evento o de la ocurrencia sea globalmente único. La instanciación completa del conjunto de datos se puede consultar en formato Turtle a través del siguiente enlace[2].

Enfoque basado en pruebas ontológicas: Este enfoque involucra preguntas de competencia (PC), que según [GF95, BFS13] son interrogativos

[1] https://mapper.obis.org/?datasetid=1cd60252-a549-4e69-b5c0-95fa0171eabb

[2] https://raw.githubusercontent.com/cenpat-gilia/BiGe-Onto/master/instances/dwca-arobis-sealsi-v1-3.ttl

orientados al usuario que nos permiten explorar la ontología. En otras palabras, son preguntas que los usuarios desean que se les responda a través de la exploración y consulta de la ontología y su base de conocimientos asociada. Para cada PC específico, desarrollamos un conjunto de casos de prueba, implementando las PC como consultas SPARQL. Elegimos SPARQL porque se puede usar para consultar el esquema RDF y el modelo OWL para filtrar personas con características específicas. Después de varias entrevistas con los expertos en dominio, los objetivos clave del modelo de ontología se especifican a continuación:

PC01: *¿Qué muestras/especies son recolectadas por un agente en particular?*
PC02: *¿Cuántas ocurrencias de una determinada especie hay?*
PC03: *¿Cuántos machos de una especie hay en el conjunto de datos?*
PC04: *¿Qué especímenes fueron recolectados en una fecha determinada?*
PC05: *¿Cuál es la clasificación taxonómica del objeto recolectado?*
PC06: *¿Dónde se ubican espacialmente los mamíferos contenidos en el conjunto de datos?*
PC07: *¿Qué ubicaciones están asociadas con una determinada especie?*
PC08: *¿Cuál es la naturaleza específica del objeto recolectado?*
PC09: *¿A qué país pertenece una ocurrencia?*
PC10: *¿A qué región marina pertenece una coordenada geográfica?*
PC11: *¿Qué ocurrencias existen dentro de un cuadro delimitador?*
PC12: *¿Qué especies coexisten en una determinada región marina?*
PC13: *¿Qué instituciones trabajan con cierta especie en Argentina?*
PC14: *¿Cuál es el ambiente asociado con una ubicación?*
PC15: *¿Qué especies de mamíferos marinos son presas o depredadores?*

Este enfoque basado en preguntas de competencia tiene las siguientes ventajas: (i) una rica expresividad semántica, que no se puede alcanzar usando solo el modelo gráfico, (ii) las inferencias (para codificar la ontología), (iii) una evaluación de la confiabilidad de la ontología propuesta, y (iv) identificación de inconsistencias. Para cada pregunta específica, desarrollamos un

conjunto de casos de prueba, implementando los preguntas como consultas SPARQL. Para una mejor legibilidad, las preguntas se han numerado y sintetizado en el Cuadro A.1 del Apéndice A. El Cuadro 5.2 incluye algunos de los casos de prueba desarrollados. Para realizar los casos de prueba, las instancias consideradas en el paso de evaluación anterior se agregaron en los archivos OWL correspondientes.

Cuadro 5.2: Casos de testeo desarrollados.

Caso	**PC**	**Entrada**	**Resultado esperado**
T001	PC01	Lewis, Mirtha	Otaria flavescens Shaw, 1800
T002	PC02	Otaria flavescens	4418
T004	PC04	28/01/1994	Otaria flavescens Shaw, 1800
T006	PC06	Mammalia	8108 positions
T009	PC09	Argentina	204761 occurrences
T013	PC013	Otaria flavescens, Argentina	ÇENPAT-CONICET"@en ANS.INST.PAT@en FPN-WCS@en CENPAT-FPN@en UNIV. AUSTRAL CHILE@en

Finalmente, después de ejecutar un caso de prueba, comparamos los resultados devueltos con los resultados esperados para determinar si el caso de prueba fue satisfactorio o no. Si los resultados coinciden, entonces la ontología *BiGe-Onto* pasa este caso de prueba. Si no es asi, debemos analizar si el problema está en el modelo conceptual, en su implementación, o incluso en la formulación/implementación de las preguntas de competencia. Para ejecutar los casos de prueba, utilizamos Protégé. La Figura 5.1 muestra un ejemplo de la ejecución del caso de prueba T002. Como podemos ver al contrastar el resultado actual con el resultado esperado, se aprobó este caso de prueba. En general, todos los casos de prueba fueron satisfactorios. Sin embargo, se detectaron varios problemas menores, la mayoría relacionados con la implementación de *BiGe-Onto*, y algunos relacionados con la implementación de los casos de prueba. Cuando se detectó un problema, hicimos los cambios necesarios y volvimos a ejecutar el caso de prueba. Un ejemplo específico donde fue necesario hacer ajustes fue en el caso de prueba T013. La especie que se buscaba era *Otaria flavescens* y el país era *Argentina*, en lugar de eso, la persona que ejecutó el caso de prueba puso el código de país *ARG* en

lugar del nombre completo del país. Este caso de prueba no mostró ningún resultado, a pesar de saber con certeza que había datos de esta especie en Argentina. Luego de este inconveniente, se llegó a un consenso para establecer reglas con respecto a los nombres de los países y los nombres científicos de la especie. Finalmente el caso de prueba se volvió a ejecutar, mostrando los resultados esperados.

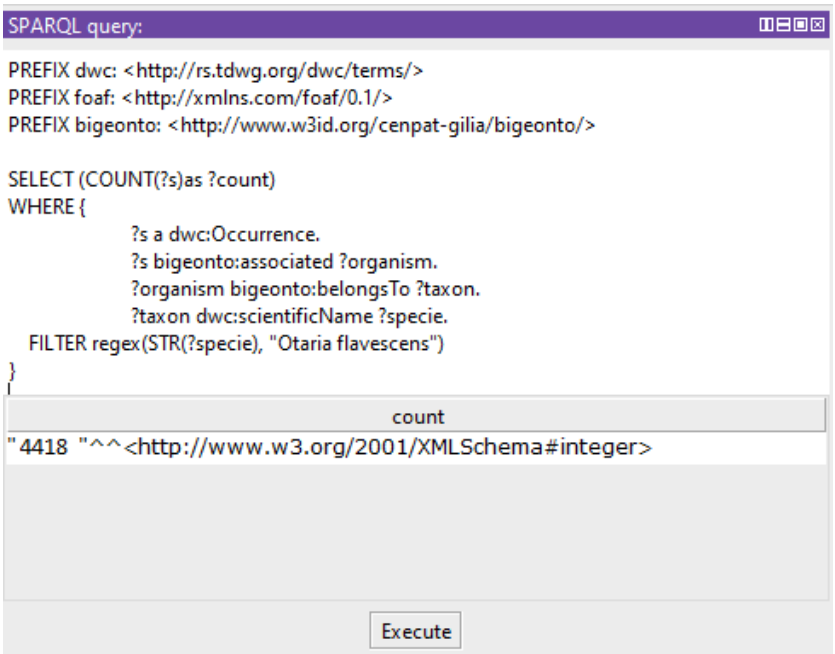

Figura 5.1: Execución en Protégé del caso de testeo T02.

5.2. Validación mediante casos de estudio

El desarrollo de los casos de estudio se inició dentro del Centro Nacional Patagónico (CENPAT)[3]. Las líneas de investigación de CENPAT incluyen biología marina, manejo de recursos acuáticos, oceanografía, paleontología y diversidad biológica, entre otros. En particular, un grupo de investigación en CENPAT llamado CESIMAR (Centro para el Estudio de Sistemas Marinos) se centra en desarrollar líneas de investigación y capacitar recursos

[3] http://www.cenpat-conicet.gob.ar/

humanos para comprender el funcionamiento de los ecosistemas marinos, y brindar apoyo científico y tecnológico para la gestión y conservación de los ecosistemas marinos. recursos Realizamos los casos de uso con tres expertos de CESIMAR. Aunque los evaluadores pertenecen en general a disciplinas relacionadas con las ciencias de la vida, están familiarizados con el uso y los beneficios de las ontologías, y son usuarios, proveedores de datos, o ambos, para GBIF u OBIS. El caso de uso se dividió en tres partes para facilitar su comprensión, la primera parte está relacionada con el modelado de la distribución de la especie, la segunda con la detección de ubicaciones sospechosas y la tercera parte muestra cómo las ocurrencias de ciertos eventos pueden relacionar las especies a las variables ambientales. De esta manera cubrimos el requisito (R1) Soporte geoespacial.

Caso de estudio 1: Distribución de especies

La distribución de especies es la manera en que un taxón biológico está dispuesto espacialmente. Los límites geográficos de la distribución de un taxón en particular son su rango, a menudo representado como áreas sombreadas en un mapa. Si bien existen servicios que pueden traducir coordenadas en una ciudad o país, aún no es posible recuperar información exacta de una región específica utilizando coordenadas geográficas de todo el mundo. Esta limitación nos llevó a pensar en formas normalizadas de definir nuestras propias áreas de interés. Para hacer esto definimos regiones usando las clases proporcionadas por la ontología GeoSPARQL. Todo lo que se requiere para vincular regiones con GeoSPARQL y así dar a sus clases una referencia geoespacial, es hacer de *bigeonto:Region* una subclase de *geo:Feature*. Por supuesto, esto puede resultar en que la clase tenga dos clases padre, pero esto es compatible con el razonamiento RDFS y OWL.

El Cuadro 5.3 muestra cómo hemos definido una región de interés llamada Golfo Nuevo (*GNPolygon*). Después de haber definido una región de estudio, queremos responder a la pregunta de interés: *¿Qué especies de mamíferos marinos se han observado dentro de Golfo Nuevo?*. Esta consulta requiere una comparación topológica entre las geometrías de las ocurrencias y las

Cuadro 5.3: Definición de la región denominada Golfo nuevo utilizando la ontología GeoSPARQL

```
bigeonto:Region a owl:Class;
 rdfs:subClassOf geo:Feature.

bigeonto:GolgoNuevo a bigeonto:Region;
 rdfs:label "Golfo Nuevo";
 geo:hasGeometry bigeonto:GNPolygon.

bigeonto:GNPolygon a geo:Polygon;
 geo:asWKT "POLYGON((-64.593493324 -42.499219410, -64.316088539
                     -42.555894203, -64.233691078 -42.650911283,
                     -64.255663734 -42.745783430, -64.143053871
                     -42.880776063, -64.340807777 -42.931070955,
                     -64.686877113 -42.888826004, -64.980761390
                     -42.792157398, -64.994494300 -42.693318880,
                     -64.857165199 -42.618581313,
                     -64.593493324 -42.499219410))"^^sf:wktLiteral.
```

geometrías de la región. Esta comparación topológica se muestra en el Listado 5.1 utilizando la propiedad *geo:sfWithin*. Tanto *Region* como la localización de la especie *sf:Point* tienen declaraciones de tipo *geo:hasGeometry* para vincularlos a sus geometrías, y luego *geo:within* para filtrar los puntos que están dentro de la región especifica como Golfo Nuevo.

Listado 5.1: Consulta para recupera ubicaciones dentro de una región especifica.

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX bigeonto: <http://www.w3id.org/cenpat-gilia/bigeonto/>
PREFIX sf: <http://www.opengis.net/ont/sf#>

SELECT ?PointWKT
WHERE {
  bigeonto:GNPolygon geo:asWKT ?PolygonWKT.
  ?point a sf:Point.
  ?point geo:asWKT ?PointWKT.
  FILTER (geof:sfWithin(?PointWKT,?PolygonWKT))
}
```

Caso de estudio 2: Detección de ubicaciones sospechosas

El segundo caso de estudio, muestra cómo utilizamos *BiGe-Onto* para realizar controles de calidad de datos, en particular, cómo detectar ubica-

ciones sospechosas. En sus requisitos de calidad de datos[4], GBIF no obliga a que un registro tenga la latitud *dwc:decimalLatitude* ni la longitud *dwc:decimalLongitude* en sus conjuntos de datos, tampoco obliga a proporcionar un código de país *dwc:countryCode*. Por esta razón, la calidad de los datos en las en las ubicaciones ha sido cuestionada por varios autores [Cha05, HOAG10, YBS+07]. La falta de información espacial precisa en los registros puede llevar a problemas, como la demarcación imprecisa de áreas protegidas para especies en peligro de extinción. En este caso de estudio, usamos la región definida en el Cuadro 5.3, para intentar encontrar ocurrencias que se observaron en Golfo Nuevo, pero que contienen algún tipo de error en el campo que describe la ubicación, por ejemplo, *Mar Argentino* en lugar de *Golfo Nuevo*. El Cuadro 5.4 muestra un ejemplo donde se puede ver un error en la propiedad *dwc:waterBody* ya que se observa la cadena de texto *Mar Argentino*, aunque esto es correcto (dado que Golfo Nuevo es parte del mar argentino) el nivel de detalle no es el correcto.

Cuadro 5.4: Definición de una ubicación utilizando GeoSPARQL

```
bigeonto:location/urncatalogcenpat-conicet-peces-p-996 a dc:location;
 dwc:decimalLatitude -47.983334;
 dwc:decimalLongitude -64.0;
 dwc:waterBody "Argentinean Sea";
 geo:hasGeometry bigeonto:point/urncatalogcenpat-conicet-peces-p-996.

bigeonto:point/urncatalogcenpat-conicet-peces-p-996 a geo:point;
 geo:asWKT "POINT(-57.516666 -41.516666 )"^^sf:wktLiteral.
```

Para detectar errores de este tipo, podemos usar nuevamente GeoSPARQL y hacer una comparación topológica entre el polígono *bigeonto:GNPolygon* y las ocurrencias que están dentro. Después de buscar todas las ocurrencias, compararemos los campos *dwc:waterBody*, *dwc:locality* y *dwc:stateProvince* para ver si son diferentes de la etiqueta *rdfs:label* "Golfo Nuevo". Si la consulta devuelve un resultado, pueden existir errores en la descripción de la ubicación. Por otro lado también podemos realizar un control de calidad de

[4] https://www.gbif.org/data-quality-requirements-occurrences

forma inversa al anterior, es decir, encontrar todas las ocurrencias que dicen pertenecer al Golfo Nuevo y obtener su latitud y longitud. Después de esto usaremos la función topológica *geo:sfDisjoint* para encontrar las ocurrencias que están fuera del polígono *bigeonto:GNPolygon*. En el Listado 5.2, se puede ver un ejemplo de esto, donde se puede usar una consulta SPARQL muy simple para realizar controles de calidad de datos.

Listado 5.2: Consulta para realizar controles de calidad en las ubicaciones.

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX bigeonto: <http://www.w3id.org/cenpat-gilia/bigeonto/>

SELECT ?PointWKT
WHERE {
 bigeonto:GNPolygon geo:asWKT ?PolWKT.
 ?loc a dc:Location.
 ?loc dwc:waterBody ?waterbody.
 ?loc geo:hasGeometry ?point.
 ?point geo:asWKT ?PointWKT
 FILTER regex(STR(?waterbody), "Golfo Nuevo")
 FILTER (geof:sfDisjoint(?PointWKT,?PolWKT ))
}
```

Caso de estudio 3: Asociando ocurrencias y temperaturas en el espacio y tiempo

Hacer coincidir las ocurrencias de especies con determinadas variables ambientales es un requisito muy común de los análisis macroecológicos, particularmente aquellos que consideran los impulsores ambientales de las distribuciones de especies, y cómo se espera que las distribuciones cambien a medida que cambia el clima. El tercer caso de estudio describe cómo podemos relacionar las ocurrencias de especies específicas con variables ambientales como la temperatura del cuerpo de agua. Para asociar las ocurrencias de una especie particular, dentro de una región como Golfo Nuevo en un rango de fechas particular, a las mediciones de la temperatura del cuerpo del agua, primero es necesario definir la región (hecha en la primera parte del caso de uso), luego recuperar las observaciones de especies para un rango de fechas dado, para esto usaremos *Time Ontology*. Y finalmente, recuperar la variable de estudio, dado que el vocabulario NERC proporciona URIs para cada una de las variables, solo necesitamos recuperar la URI especifica de

la temperatura del cuerpo de agua, la cual es http://vocab.nerc.ac.uk/collection/P01/current/TEMP/. El Cuadro 5.5 muestra un fragmento en RDF donde se puede ver que la medición *measurement1* corresponde a la medicion de la temperatura cuyo identificador es *TEMP*. El valor asociado *dwc:MeasurementValue* es un entero igual a seis y la unidad de medida *dwc:MeasurementUnitID* corresponde a grados centígrados, también identificada por una URI definida en NERC, cuyo identificador es *UPAA*.

Cuadro 5.5: Definición en RDF de la variable ambiental *temperatura del cuerpo de agua (TEMP)*

```
bigeonto:ExtendedMeasurementOrFact a owl:Class.

bigeonto:measurement1 rdf:type bigeonto:ExtendedMeasurementOrFact;
rdfs:label "Medicion de temperatura de la columna de agua";
dwc:MeasurementTypeID http://vocab.nerc.ac.uk/collection/P02/current/TEMP/;
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/urncatalog-conicet-peces-p-331.

bigeonto:occurrence/urncatalog-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/urncatalog-conicet-peces-p-331 .
```

Finalmente, en el Listado 5.3 definimos una consulta SPARQL que asocia las ocurrencias de una especie marina (en nuestro caso un pez cuyo nombre científico es *Merluccius hubbsi*) con la temperatura del cuerpo del agua en una región marina específica. En primer lugar, definimos la región marina llamada *Golfo de San Matias*, de tipo (*geo:Polygon*), en segundo lugar recuperamos las observaciones de las especies definidas como puntos usando (*geo:point*). Dado que GeoSPARQL permite realizar operaciones espaciales,

podemos consultar si un punto está contenido dentro de un polígono usando la función (*geof:sfWithin*). Finalmente recuperamos la variable ambiental medida (también georreferenciada). NERC proporciona URIs para cada una de las variables, por lo que solo necesitamos recuperar la URI para la temperatura *TEMP*.

Listado 5.3: Consulta para asociar ocurrencias de una especie con variables ambientales en una región particular.

```
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"))
}
```

En esta sección hemos utilizado la ontología GeoSPARQL para definir regiones y realizar operaciones geospaciales sobre las mismas, mostrando como se pueden integrar observaciones de especies de GBIF con variables ambientales de OBIS, de esta manera cubrimos uno de los requisitos mas importantes de esta investigación (R1) Soporte geoespacial.

5.3. Comparando *BiGe-Onto* con otros Sistemas/Ontologías

Para proporcionar la base de nuestro modelo conceptual (Sección 4.3), realizamos una revisión de la literatura en busca de ontologías de dominio compartido. A pesar de nuestra extensa investigación, encontramos trabajos de investigación que se basan únicamente en ontologías de Biodiversidad u Oceanografía, pero no en ambas. Seguimos los criterios propuestos en [dG11] para evaluar ontologías, a saber (i) tener una buena cobertura de dominio (Biodiversidad/Ocenografía o ambas); (ii) implementar un estándar internacional como GeoSPARQL; (iii) ser formalmente rigurosa, teniendo en cuenta si la ontología presenta axiomas definidos en algún lenguaje formal; (iv) ser modular y diseñada siguiendo determinados principios; (v) seguir un metodología de evaluación para probar la ontología; (vi) reutilizar ontologías fundacionales que verifiquen la confiabilidad de las mismas; (vii) implementar un método de ingeniería ontológica; y (viii) publicar información en formato RDF para su accesibilidad y reutilización. Además, a nivel de sistema, definimos cuatro puntos que son importantes al evaluar la calidad, estos son: (i) proporcionan un punto final SPARQL, de esta manera diferentes aplicaciones pueden reutilizar los datos; (ii) proporcionar una interfaz web, para que el usuario no experto pueda interpretar la información fácilmente y con un conocimiento mínimo de los lenguajes de programación; (iii) proporcionar una API REST para recuperar los datos; y (iv) proporcionar documentación clara con ejemplos concretos de uso. Es importante destacar que de manera individual o de manera conjunta estos puntos contribuyen a alcanzar cada uno de los requisitos definidos en la sección 3.3: **(R1)** Soporte geoespacial, **(R2)** Interoperabilidad con otros sistemas, **(R3)** Interfaz visual, **(R4)**, Proveer un punto de acceso, **(R5)** Metodología formal y **(R6)** Cobertura de ambos dominios.

Dentro de las ontologías y sistemas relevados, podemos mencionar las siguientes:

[WDG+14] describe la Ontología de Colecciones Biológicas (BCO). Esta ontología aborda específicamente la necesidad de integrar los registros de ocurrencia de especies con datos moleculares, ecológicos y de fenotipo. Los estudios a gran escala realizados por científicos de todo el mundo han proporcionado casos de uso sólido para guiar el desarrollo de esta ontología. Aunque BCO es ampliamente aceptado por la comunidad de la biodiversidad y las ciencias de la vida en general, no propone el uso de estándares para consultar datos geoespaciales como GeoSPARQL, lo que sería muy útil en estudios relacionados con biogeografía.

[SDC+14] describe un software para convertir datos de Biodiversidad en formatos tabulares estándar, como DwC-A, en representaciones RDF. En este caso, se describe una ontología llamada *BiSciCol* utilizando los términos del DwC, pero no hay ejemplos de inferencias o consultas que puedan ayudar a los investigadores interesados en integrar diferentes bases de datos. Por otro lado, no describe una metodología formal para el desarrollo de la ontología.

[BW15] presentan Darwin-SW (DSW) una ontologia y un vocabulario RDF diseñado para complementar el estándar Darwin Core. DSW se basa en un modelo derivado de un consenso de la comunidad sobre las relaciones entre las clases principales de Darwin Core. DSW crea nuevas clases para acomodar aspectos importantes de su modelo que actualmente no forman parte de Darwin Core. DSW usa OWL para hacer afirmaciones sobre las clases en su modelo y para definir las propiedades objeto que se usan para vincular instancias de esas clases. Un objetivo en la creación de DSW era facilitar un marcado consistente de los datos de Biodiversidad para que los grafos RDF creados por diferentes proveedores pudieran fusionarse fácilmente. Al igual que el anterior, este trabajo no describe una metodología formal para el desarrollo de la ontología ni utiliza estándares como GeoSPARLQ para la gestión de datos geoespaciales.

[SSF+18] presenta una ontología llamada *OpenBiodiv-O* que introduce clases, propiedades y axiomas en los dominios de la publicación académica de Biodiversidad y la taxonomía biológica, y los alinea con varias ontologías

de dominio importantes (incluyendo Darwin-SW [BW15] y EnvO [BPL+16]). El objetivo principal de este trabajo es reducir la brecha ontológica entre la publicación de Biodiversidad académica y taxonomía biológica y permite la creación de un conjunto de datos abiertos vinculados de información de Biodiversidad. Como todos los trabajos anteriores, no se describe una metodología formal para el desarrollo de la ontología ni utiliza estándares relevantes como GeoSPARLQ para la gestión de la información geoespacial.

El Cuadro 5.6 resume las diferencias entre *BiGe-Onto* y las propuestas disponibles en la literatura. Para evaluar si la ontología se desarrolló de una manera formalmente rigurosa o no, tenemos en cuenta si la ontología también presenta axiomas definidos en algún lenguaje formal. Con respecto a la revisión realizada, ninguna de las contribuciones en la literatura discutió cómo se evaluaron las ontologías propuestas. Con respecto a la reutilización de ontologías fundacionales solamente, [WDG+14] describen el uso de Ontología Formal Básica (BFO). No usar ontologías fundacionales puede considerarse un problema debido a las importantes distinciones que se hacen en ontologías formales. Además, la falta de hallazgos verdaderamente ontológicos pone en peligro la veracidad de esas ontologías. La fácil disponibilidad de datos de Biodiversidad/Biogeografía como RDF permite a los investigadores combinar datos de diferentes fuentes y analizarlos con poderosas herramientas de consulta de datos vinculados. De la revisión de la literatura, llegamos a la conclusión de que [SSF+18] es la única propuesta que publica sus datos en RDF. Otros desafíos, que abordamos en *BiGe-Onto*, son la capacidad de expresar consultas complejas, es decir, consultas que utilizan relaciones espaciales ("ocurrencias dentro de un área protegida"). En contraste, [WDG+14, SSF+18, BW15, SDC+14] pueden responder consultas SPARQL, pero no pueden manejar consultas que tratan con conceptos espaciales.

Cuadro 5.6: Comparativa de *BiGe-Onto* frente a los trabajos relevados. **OK** indica que la característica es soportada, **parcialmente** indica un soporte parcial, y "**-**" indica que la característica no es soportada o no se especifica. Cada una de las características contribuye a cumplir con los requisitos previamente definidos.

Req.	**Característica**	**BiGe-Onto**	**Darwin-SW**	**OpenBiodiv-O**	**BiSciCol**	**BCO**
Nivel Ontología						
R1	(ii) Soporte Geoespacial	OK	-	-	-	-
R6	(i) Cobertura del dominio	Biodiversidad Biogeografía	Biodiversidad	Biodiversidad	Biodiversidad	Biodiversidad
R5	(iii) Axiomas	OWL	OWL	OWL	OWL	OWL
R5	(iv) Modularidad	OK	parcialmente	parcialmente	parcialmente	OK
R5	(v) Método de evaluación	[BGM05, VG06]	-	-	-	Expertos de dominio
R5	(vi) Ontologias fundacionales	UFO	-	-	-	BFO
R5	(vii) Método de ingeniería	SABiO	-	-	-	-
R2,R4	(viii) Disponibilidad como RDF	OK	OK	OK	OK	OK
Nivel Sistema						
R4	(i) Interfaz SPARQL	OK	-	OK	-	-
R3	(ii) Interfaz visual	parcialmente	-	OK	OK	-
R2	(iii) API REST	OK	-	OK	-	-
R3	(iv) Documentación/ejemplos	OK	OK	OK	parcialmente	OK

En este capitulo presentamos los resultados de la evaluación de *BiGe-Onto* desde dos perspectivas. La primera de ellas tiene que ver con la validación del modelo conceptual mediante casos de testeo. La segunda perspectiva tiene que ver con la validación mediante casos de estudio donde se demuestra la utilidad de *BiGe-Onto* para el manejo integrado de información.

Finalmente comparamos *BiGe-Onto* con otras ontologías y sistemas, como lo muestra el Cuadro 5.6, la principal característica distintiva de *BiGe-Onto* cuando se compara con otras ontologías es que, según nuestro conocimiento, es el primer esfuerzo que cubre aspectos que son comunes en diferentes áreas de las ciencias de la vida. Por ejemplo, *BiGe-Onto* utiliza GeoSPARQL para proporcionar razonamiento espacial, como relaciones topológicas, o superposiciones entre entidades espaciales. Además, *BiGe-Onto* se desarrolló siguiendo la metodología SABiO, que es un método definido formalmente y ampliamente utilizado en diversos desarrollos de ontologías, esto proporciona solidez en la validación. En el siguiente capitulo presentamos las conclusiones que se desprenden de esta investigación como asi también los trabajos futuros o cuestiones sin resolver aún en el marco de esta tesis.

Capítulo 6

Conclusiones y Trabajos Futuros

Índice

Este capítulo expone las ideas finales de esta tesis, las conclusiones obtenidas en todos los experimentos realizados y las líneas de trabajo futuras que planeamos explorar a partir de los últimos trabajos.

La cantidad de datos en la Web ha aumentado enormemente en las últimas dos décadas, específicamente, en el dominio de las Ciencias de la Vida. Esta cantidad de información hace necesario incluir semántica para que sea procesada por las máquinas. La tecnología de los Datos Vinculados surgió en 2008 y puede definirse como un conjunto de prácticas recomendadas para compartir, exponer y conectar información, datos y conocimientos mediante el uso de los estándares RDF y OWL. Esta tecnología, apoyada por la comunidad W3C, se ha aplicado a diferentes categorías de datos, como Gobierno, Publicaciones, Web social, Dominio cruzado, Geográfico, Medios de comunicación, Generado por el usuario y Ciencias de la vida.

Como expone [GW06] existen diversos problemas en el manejo integrado

de información en el campo de las ciencias de la vida y en particular en la Biodiversidad y la Oceanografía, en este sentido, solo unos pocos esfuerzos han sido presentados y analizados, pero ninguno de ellos cubre todos los requisitos definidos en el Capítulo 3. Por este motivo, en esta tesis presentamos el desarrollo de un sistema basado en ontologías llamado *BiGe-Onto* para administrar la información asociada con Biodiversidad/Biogeografía marina, utilizando estándares como Darwin Core y GeoSPARQL.

En una primera etapa, se desarrollo un marco para publicar datos de biodiversidad de la Patagonia argentina como datos abiertos enlazados [ZBF]. Estos conjuntos de datos contenían información de especies biológicas (mamíferos, plantas, parásitos, entre otros) que han sido recopilados por investigadores de Centro Nacional Patagónico, y que inicialmente estabas disponibles como archivos DwC-A. Introducimos y detallamos un proceso de transformación y explicamos cómo acceder y explotarlos, promoviendo la integración con otros repositorios como DBPedia. Como paso siguiente expandimos este trabajo desarrollando una ontología para integrar información Biológica y Biogeografía [ZBF17], de esta primera obtuvimos una importante feedback respecto de la usabilidad de la misma por parte de los usuarios y que mejoras deberíamos hacer a este modelo. Posteriormente presentamos los pasos iniciales en la creación de un conjunto de datos oceanográficos vinculados utilizando información de las campañas oceanográficas de la iniciativa Argentina Pampa Azul [ZRF+18]. Para lograr consistencia, capacidad de descubrimiento y hacer que los conjuntos de datos sean legibles por máquinas y humanos, usamos diferentes vocabularios controlados, entre ellos NERC, GeoSPARQL. Reutilizamos el patrón de diseño ontológico para cruceros oceanográficos. Además, complementamos la información de Pampa Azul con el conjunto de datos vinculados oceanográficos R2R. A partir de estas contribuciones anteriores, se planteo el desarrollo de un grafo de conocimiento oceanográfico [ZRB+19] donde entidades tales como publicaciones científicas, personas, lugares, especímenes, variables ambientales e instituciones forman parte de un único espacio de conocimiento compartido, en este articulo describimos el proceso de modelado y publicación, y los usos actuales y futuros del

conjunto de datos. Como corolarios del trabajo realizado se desprenden dos contribuciones, la primera tiene que ver con la calidad de los datos georeferenciados [ZLFD18], donde se utilizan datos abiertos enlazados y GeoSPARQL para detectar errores en la georeferencias de manera semiautomática. Y la segunda [ZL16] con los problemas de interoperabilidad que afronta el sistema de observación global de los océanos, proponiendo buenas practicas y recomendaciones para abordarlos.

Finalmente, se presentó una implementación concreta para contribuir a solucionar el problema del manejo integrado de información de Biodiversidad y Biogeografía marina, esta implementación se denomino *BiGe-Onto* [ZBF+19]. Un sistema basado en ontologías para el manejo integrado de información. Compuesto de (i) la arquitectura; (ii) un modelo conceptual especificado en OntoUML; (iii) una versión operacional implementada en OWL 2; y (iv) un conjunto de datos integrado para su explotación a través de un punto final SPARQL. La primer versión oficial de *BiGe-Onto* se encuentra en linea en `http://crowd.fi.uncoma.edu.ar/cenpat-gilia/bigeonto/` y el conjunto de datos enlazados es acesible a través de `https://doi.org/10.5281/zenodo.3235548`.

La evaluación realizada a *BiGe-Onto* muestra que cumple con el propósito inicial de esta tesis: *gestionar de forma integrada la información de Biodiversidad y Biogeografía marina*, y nos permite responder las preguntas científicas que motivaron nuestra propuesta. Como lo muestra la Tabla 5.6, la principal característica distintiva de *BiGe-Onto* cuando se compara con otras ontologías es que, según nuestro conocimiento, es el primer intento que cubre aspectos que son comunes en diferentes áreas de las ciencias de la vida. Por ejemplo, *BiGe-Onto* utiliza GeoSPARQL para permitir incluir una forma estándar de proporcionar razonamiento espacial, como relaciones topológicas, o superposiciones entre entidades espaciales. Además, *BiGe-Onto* se desarrolló siguiendo la metodología SABiO, que es un método bien establecido, utilizado en diversos esfuerzos de desarrollo de ontología, lo que proporciona un medio de validación de solidez. La versión actual no tiene antipatrones, lo que puede interpretarse en el sentido de que la ontología se encuentra en

una versión estable que puede evolucionar agregando conceptos que pueden ampliar el alcance de la ontología si es necesario. Con respecto a la evaluación del conjunto de datos *BiGe-Onto*, proporcionamos un punto de acceso SPARQL para que otras aplicaciones puedan consumir y reutilizar nuestra información. Presentamos un conjunto de consultas SPARQL definidas por el usuario, para demostrar cómo podemos gestionar la información de OBIS y GBIF.

6.1. Trabajos Publicados

Las publicaciones que surgieron de esta tesis se pueden organizar de la siguiente manera según su tipo de publicación:

Artículos publicados en revistas:

1. **Zárate, M.**, Braun, G., Fillottrani, P., Delrieux, & C.Lewis, M. (2019). BiGe-Onto: An Ontology-Based System for Managing Biodiversity and Biogeography Data. Applied Ontology journal, Accepted.

2. Snowden, Derrick P, Tsontos, Vardis, Handegard, Nils Olav, **Zárate, M**, O'Brien, Kevin M, Casey, Kenneth S, Smith, Neville, Sagen, Helge, Bailey, Kathleen, Lewis, Mirtha and others (2019). Data Interoperability Between Elements of the Global Ocean Observing System. Frontiers in Marine Science, 6, 442. DOI: 10.3389/fmars.2019.00442

3. **Zárate, M.**, Buckle, C., Mazzanti, R., Samec, G. (2019). Improving Open Science Using Linked Open Data: CONICET Digital Use Case. Journal of Computer Science & Technology, 19, 45–54.
   DOI: 10.24215/16666038.19.e05

4. **Zárate, M.**, Lewis, M. (2016). Estimate of the Anesthesia Stage in Southern Elephant Seals using WEKA Data Mining Tool. International Journal of Applied Information Systems (IJAIS), 11(4):48-52, DOI: 10.5120/ijais2016451603

Capítulos en libros:

5. **Zárate, M.**, Rosales, P., Braun, G., Lewis, M., Fillottrani, P., & Delrieux, C. (2019). OceanGraph: Some Initial Steps Toward a Oceanographic Knowledge Graph. In: Villazón-Terrazas B., Hidalgo-Delgado Y. (eds) Knowledge Graphs and Semantic Web. KGSWC 2019. Communications in Computer and Information Science, vol 1029. Springer, Cham. DOI: 10.1007/978-3-030-21395-4_3

Artículos publicados en conferencias internacionales:

6. **Zárate, M.**, Rosales, P., Fillottrani, P., Delrieux, C., & Lewis, M. (2018). Oceanographic Data Management: Towards the Publishing of Argentine Oceanographic Campaigns as Linked Data, 12^{th} Alberto Mendelzon International Workshop on Foundations of Data Management. Cali, Colombia: URL: ceur-ws.org/Vol-2100/paper-20

7. **Zárate, M.**, Lewis, M., Fillottrani, P. & Delrieux, C. (2018). Improving the Quality of Biodiversity Data Through Semantic Web Standards, 10^{th} International Conference on Ecological Informatics (pp. 245–246). Jena, Germany. URL: Proceedings

8. Mazzanti, R., **Zárate, M.**, Samec, G., Buckle, C. (2018). Integración de repositorios semánticos: un camino hacia los datos abiertos enlazados , 8^{th} BIREDIAL ISTEC 2018 Conference, Lima, Peru. URL: Proceedings

9. **Zárate, M.**, Braun, G., & Fillottrani, P. (2017). Adding Biodiversity Datasets from Argentinian Patagonia to the Web of Data, 2^{nd} International Workshop on Semantics for Biodiversity co-located with 16^{th} International Semantic Web Conference 2017. Vienna, Austria: URL: ceur-ws.org/Vol-1933/paper-6

10. **Zárate, M.**, Buccella, A., & Fillottrani, P. (2017). BioOnto: Towards an Integration of Biological and Biogeographic Data, 16^{th} Joint Ontology Workshops 2017, Bozen-Bolzano, Italy. URL: ceur-ws.org/Vol-2050/ODLS-paper-6

11. **Zárate, M.**, Braun, G., & Almonacid, S. (2017). Transforming Biodiversity Data in Linked Data, 19^{th} Workshop of Researchers in Computer Science, Buenos Aires, Argentina. URL: ISBN-978-987-42-5143-5

Artículos publicados en conferencias nacionales:

12. Samec, G., Diez Maria E., **Zárate, M.**, Buckle, C., Lima, J., Jaramillo, R., Sanchez A., & Mazzanti, R., (2019). Aplicaciones Informáticas para el Estudio de Biodiversidad de Poliquetos Espiónidos en los Golfos Nordpatagónicos, 21^{th} Workshop of Researchers in Computer Science, San Juan, Argentina. ISBN 978-987-3984-85-3

13. Mazzanti, R., **Zárate, M.**, Samec, G., Buckle, C. (2018). Infraestructura de Acceso a Datos Primarios con Aporte de Semántica en Repositorios Digitales , 20^{th} Workshop of Researchers in Computer Science, Buenos Aires, Argentina. ISBN 978-987-3619-27-4

14. **Zárate, M.**, Braun, G., & Almonacid, S. (2017). Transforming Biodiversity Data in Linked Data, 19^{th} Workshop of Researchers in Computer Science, Buenos Aires, Argentina. URL: ISBN-978-987-42-5143-5

15. **Zárate, M.**, Lewis, M. (2016). Estimación del Plano Anestésico en Elefante Marinos del Sur Utilizando Técnicas de Machine Learning, 12^{th} Argentine Congress of Computer Science, San Luis, Argentina. URL: sedici.unlp.edu.ar/handle/10915/56749

16. Delrieux, C., Barry, D., Stickar, R. Mazzanti, R. Buckle, C. Cura, R. & **Zárate, M.** (2015). Classification of information in Big Data through the use of artificial intelligence techniques and social network analysis, 17^{th} Workshop of Researchers in Computer Science, Salta, Argentina, ISBN-978-987-42-5143-5

6.2. Trabajo Futuro

A partir de la investigación presentada en esta tesis se abren las siguientes líneas de investigación.

En primer lugar, estamos planificando como línea de investigación posdoctoral continuar explorando más a fondo cómo se pueden utilizar las ontologías y los datos enlazados para gestionar el conocimiento en el contexto de la Oceanografía. En particular como se puede realizar la gestión integrada de datos científicos multidisciplinares y visualización proveniente de datos de campañas oceanográficas, de repositorios de Biodiversidad y de datos ambientales. En una primera etapa será acotado al Golfo San Jorge[1] dado que se cuenta con datos provenientes de las campañas realizadas por el grupo de trabajo del Golfo San Jorge, perteneciente a la Iniciativa Pampa Azul, y luego puede ser escalable a otros espacios marinos donde se cuente con información de campañas oceanográficas así como estaciones fijas con sensores ambientales remotos.

Debido a que uno de los objetivos que se definió en un principio era utilizar *BiGe-Onto* para realizar controles de calidad en los datos y no fue posible alcanzarlo, tenemos la intención de realizar una evaluación más consistente de *BiGe-Onto* para usarlo en colaboración con el nodo argentino de OBIS[2] para detectar y validar conjuntos de datos. Esperamos que *BiGe-Onto* pueda ser utilizado por especialistas interesados en estas problemática y que esté sujeto a la adición de módulos especializados como un efecto de su explotación por diferentes comunidades en diferentes áreas.

[1] http://www.pampazul.gob.ar/areas-prioritarias/golfo-san-jorge/
[2] https://obis.org/node/464a96d8-c17e-4bbb-b6b8-778e1fb687c4

Apéndice A

Preguntas de competencia implementas en SPARQL

En este apéndice, desarrollamos las preguntas de competencia formuladas en la Sección 5.1.

PC01: ¿Qué muestras/especies son recolectadas por un agente en particular?
PC02: ¿Cuántas ocurrencias de una determinada especie hay?
PC03: ¿Cuántos machos de una especie hay en el conjunto de datos?
PC04: ¿Qué especímenes fueron recolectados en una fecha determinada?
PC05: ¿Cuál es la clasificación taxonómica del objeto recolectado?
PC06: ¿Dónde se ubican espacialmente los mamíferos contenidos en el conjunto de datos?
PC07: ¿Qué ubicaciones están asociadas con una determinada especie?
PC08: ¿Cuál es la naturaleza específica del objeto recolectado?
PC09: ¿A qué país pertenece una ocurrencia?
PC10: ¿A qué región marina pertenece una coordenada geográfica?
PC11: ¿Qué ocurrencias existen dentro de un cuadro delimitador?
PC12: ¿Qué especies coexisten en una determinada región marina?
PC13: ¿Qué instituciones trabajan con cierta especie en Argentina?

Cuadro A.1: Preguntas de competencia implementadas en SPARQL.

ID	**Consulta SPARQL**

PC01	SELECT ?specimen ?name
	WHERE {
	?s a dwc:Occurrence.
	?s dwciri:recordedBy ?person.
	?person foaf:name ?name.
	?s bigeonto:associated ?organism.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:scientificName ?specimen.
	FILTER regex(STR(?name), ‘‘Agent Name’’)
	}

PC02	SELECT (COUNT(?s)as ?count)
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:associated ?organism.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:scientificName ?specie.
	FILTER regex(STR(?specie),‘‘scientific name’’)
	}

PC03	SELECT ?name (COUNT(?sex) as ?count)
	WHERE {
	?s a dwc:Occurrence.
	?s dwc:sex ?sex.
	?s bigeonto:associated ?organism.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:scientificName ?name.
	FILTER regex(STR(?sex), ‘‘male’’)
	}
	GROUP BY ?name

Tabla A.1 continua desde la página anterior

PC04	SELECT ?specimen ?date
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:associated ?organism.
	?s bigeonto:has_event ?event.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:scientificName ?specimen.
	?event dwc:eventDate ?date
	FILTER (?date = ‘‘put date’’^^xsd:date)
	}

PC05	SELECT ?taxon ?class ?family ?genus
	?kingdom ?order ?phylum ?name
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:associated ?organism.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:class ?class.
	?taxon dwc:family ?family.
	?taxon dwc:genus ?genus.
	?taxon dwc:kingdom ?kingdom.
	?taxon dwc:order ?order.
	?taxon dwc:phylum ?phylum.
	?taxon dwc:scientificName ?name
	}

PC06	SELECT ?name ?lat ?long
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:associated ?organism.
	?s bigeonto:has_event ?event.
	?event bigeonto:has_location ?location.
	?location geo-pos:lat ?lat.
	?location geo-pos:long ?long.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:class ?class.
	?taxon dwc:scientificName ?name
	FILTER regex(STR(?class), ‘‘Mammalia’’)
	}

Tabla A.1 continua desde la página anterior

PC07	SELECT ?name ?lat ?long
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:associated ?organism.
	?s bigeonto:has_event ?event.
	?event bigeonto:has_location ?location.
	?location geo-pos:lat ?lat.
	?location geo-pos:long ?long.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:scientificName ?name
	FILTER regex(STR(?name), ‘‘Scientific name’’)
	}

PC08	SELECT ?specimen ?type (COUNT(?type) as ?count)
	WHERE {
	?s a dwc:Occurrence.
	?s dwc:basisOfRecord ?type.
	?s bigeonto:associated ?organism.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:scientificName ?specimen
	}
	GROUP BY ?specimen ?type
	ORDER BY ?specimen

PC09	SELECT (COUNT(?s) AS ?count)
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:has_event ?event.
	?event bigeonto:has_location ?loc.
	?loc dwc:country ?code.
	FILTER regex(STR(?code), ‘‘country’’)
	}

PC10	SELECT ?region ?lat ?long
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:has_event ?event.
	?event bigeonto:has_location ?loc.
	?loc geo-pos:lat ?lat.
	?loc geo-pos:long ?long.
	?loc bigeonto:into_eez ?region
	}

Tabla A.1 continua desde la página anterior

PC11	SELECT ?s
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:has_event ?event.
	?event bigeonto:has_location ?location.
	?loc geo-pos:lat ?lat.
	?loc geo-pos:long ?long.
	FILTER(?long >MAX-LON && ?long <MIN-LONG &&
	?lat >MAX-LAT && ?lat <MIN-LAT).
	}

PC12	SELECT ?s_name
	WHERE {
	?s a dwc:Occurrence.
	?s bigeonto:has_taxon ?taxon.
	?taxon dwc:scientificName ?s_name.
	?taxon dwc:class ?class.
	?s bigeonto:has_event ?event.
	?event bigeonto:has_location ?loc.
	?loc bigeonto:into_eez ?region
	FILTER regex(STR(?region), ‘‘region name’’)
	}

PC13	SELECT DISTINCT ?institution_name
	WHERE {
	?s a dwc:Occurrence.
	?s dwciri:inDataset ?dataset.
	?dataset dwc:institutionCode ?institution.
	?institution rdfs:label ?institution_name.
	?s bigeonto:has_event ?event.
	?s bigeonto:associated ?organism.
	?organism bigeonto:belongsTo ?taxon.
	?taxon dwc:scientificName ?specie.
	?event bigeonto:has_location ?loc.
	?loc dwc:country ?code.
	FILTER regex(STR(?specie), ‘‘Scientific name’’)
	FILTER regex(STR(?code), ‘‘Argentina’’)
	}

Bibliografía

[ABC19] Access to biological collections data task group. 2007. access to biological collection data (abcd), version 2.06. biodiversity information standards (tdwg). http://www.tdwg.org/standards/115, 2019. [Online; accessed 29-May-2019].

[ACHZ11] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing linked datasets with the void vocabulary. 2011.

[ACS+13] Robert Arko, Cynthia Chandler, Karen Stocks, Shawn Smith, Paul Clark, Adam Shepherd, Carla Moore, and Stace Beaulieu. Rolling deck to repository (r2r): Collaborative development of linked data for oceanographic research. In *EGU General Assembly Conference Abstracts*, volume 15, 2013.

[ADA16] Hiteshwar Kumar Azad, Akshay Deepak, and Kumar Abhishek. Linked open data search engine. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, ICTCS '16, pages 17:1–17:5, New York, NY, USA, 2016. ACM.

[AGP10] Marcelo Arenas, Claudio Gutierrez, and Jorge Pérez. On the semantics of sparql. In *Semantic Web Information Management*, pages 281–307. Springer, 2010.

[ASB09] SFCD Araújo, Daniel Schwabe, and Simone Barbosa. Experimenting with explorator: a direct manipulation generic rdf browser and querying tool. *Visual Interfaces to the Social and the Semantic Web (VISSW 2009), Sanibel Island, Florida*, 2009.

[Baa09] Franz Baader. *Description Logics*, pages 1–39. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[BCH+07] Chris Bizer, Richard Cyganiak, Tom Heath, et al. How to publish linked data on the web. 2007.

[BCO] The biological and chemical oceanography data management office. https://www.bco-dmo.org/. [Online; accessed 20-May-2019].

[BdSS+17] Pedro Paulo F. Barcelos, Victor Amorim dos Santos, Freddy Brasileiro Silva, Maxwell E. Monteiro, and Anilton Salles Garcia. An automated transformation from ontouml to OWL and SWRL. In *Proceedings of the 2nd International Workshop on Semantics for Biodiversity co-located with 16th International Semantic Web Conference (ISWC 2017)*, 2017.

[BFS13] Camila Bezerra, Fred Freitas, and Filipe Santana. Evaluating ontologies with competency questions. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, volume 3, pages 284–285. IEEE, 2013.

[BG09] Alessander Botti Benevides and Giancarlo Guizzardi. A model-based tool for conceptual modeling and domain ontology engineering in ontouml. *Enterprise Information Systems*, pages 528–538, 2009.

[BGM05] Janez Brank, Marko Grobelnik, and Dunja Mladenić. A survey of ontology evaluation techniques. 2005.

[BL06] Tim Berners-Lee. Linked data. https://www.w3.org/DesignIssues/LinkedData.html, 2006. [Online; accessed 8-May-2019].

[BLCC+06] Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd international semantic web user interaction workshop*, volume 2006, page 159. Citeseer, 2006.

[BLHL+01] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[BM14] Dan Brickley and Libby Miller. Foaf vocabulary specification 0.99. http://xmlns.com/foaf/spec/, 2014. Online; accessed 29 March 2019.

[BMS+13] Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J. Mungall, and Suzanna E. Lewis. The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, 4(1):43, Dec 2013.

[BNT+08] 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*, 41(5):706–716, 2008.

[BPL+16] 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.

[BW15] Steven J Baskauf and Campbell O Webb. Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. 2015.

[CGQ08] Gong Cheng, Weiyi Ge, and Yuzhong Qu. Falcons: searching and browsing entities on the semantic web. In *Proceedings of the 17th international conference on World Wide Web*, pages 1101–1102. ACM, 2008.

[CGS+13] CL Chandler, RC Groman, A Shepherd, MD Allison, D Kinkade, S Rauch, PH Wiebe, and DM Glover. Using controlled vocabularies and semantics to improve ocean data discovery. In *AGU Fall Meeting Abstracts*, 2013.

[Cha05] Arthur D Chapman. *Principles of data quality*. GBIF, 2005.

[CKA+18] Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. The geolink knowledge graph. *Big Earth Data*, 2018.

[dAF14] Ricardo de Almeida Falbo. Sabio: Systematic approach for building ontologies. In *Proceedings of the 1st Joint Workshop ONTO.COM / ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering co-located with 8th International Conference on Formal Ontology in Information Systems.*, 2014.

[Dat] Data set rdf dumps. https://www.w3.org/wiki/DataSetRDFDumps. [Online; accessed 8-May-2019].

[DBP19] Dbpedia endpoint. https://dbpedia.org/sparql, 2019. [Online; accessed 15-May-2019].

[DFJ+04] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 652–659. ACM, 2004.

[dG11] Mathieu d'Aquin and Aldo Gangemi. Is there beauty in ontologies? *Applied Ontology*, 6(3):165–175, 2011.

[DIS07] Disco - hyperdata browser. http://wifo5-03.informatik.uni-mannheim.de/bizer/ng4j/disco/, 2007. [Online; accessed 20-May-2019].

[Eck95] Wayne W Eckerson. Three tier client/server architecture: Achieving scalability, performance, and efficiency in client server applications. *Open Information Systems*, 1995.

[GBI19] Gbif the global biodiversity information facility. https://www.gbif.org/, 2019. [Online; accessed 27-May-2019].

[Geo07] Geoowl. http://www.w3.org/2005/Incubator/geo/XGR-geo-20071023/W3C_XGR_Geo_files/geo_2007.owl, 2007. [Online; accessed 24-May-2019].

[Geo11] Geordf: An rdf compatible profile for geo information (points, lines and polygons). https://www.w3.org/wiki/GeoRDF, 2011. [Online; accessed 24-May-2019].

[GF95] Michael Grüninger and Mark S Fox. The role of competency questions in enterprise engineering. In *Benchmarking—Theory and practice*, pages 22–31. Springer, 1995.

[GHM+08] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. Owl 2: The next step for owl. *Journal of Web Semantics*, 6(4):309 – 322, 2008. Semantic Web Challenge 2006/2007.

[GHM+14] Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stoilos, and Zhe Wang. Hermit: an owl 2 reasoner. *Journal of Automated Reasoning*, 53(3):245–269, 2014.

[GRP+16] Ilse R Geijzendorffer, Eugenie C Regan, Henrique M Pereira, Lluis Brotons, Neil Brummitt, Yoni Gavish, Peter Haase,

Corinne S Martin, Jean-Baptiste Mihoub, Cristina Secades, et al. Bridging the gap between biodiversity data and policy reporting needs: An essential biodiversity variables perspective. *Journal of Applied Ecology*, 53(5):1341–1350, 2016.

[Gui05] Giancarlo Guizzardi. *Ontological foundations for structural conceptual models*. CTIT, Centre for Telematics and Information Technology, 2005.

[GW06] Benjamin M Good and Mark D Wilkinson. The life sciences semantic web is full of creeps! *Briefings in bioinformatics*, 7(3):275–286, 2006.

[GW10] Giancarlo Guizzardi and Gerd Wagner. *Using the Unified Foundational Ontology (UFO) as a Foundation for General Conceptual Modeling Languages*. Springer Netherlands, 2010.

[Hau09] Michael Hausenblas. Linked data applications. *First Community Draft, DERI*, 2009.

[HB11] Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[HDHC14] Conor Hayes, Stefan Decker, Benjamin Heitmann, and Richard Cyganiak. Architecture of linked data applications. In *Linked Data Management: Principles and Techniques*. CRC Press (Taylor & Francis), 2014.

[HG04] Jonathan Hayes and Claudio Gutierrez. Bipartite graphs as intermediate model for rdf. In *International Semantic Web Conference*, pages 47–61. Springer, 2004.

[HHSS12] Peter Haase, Christian Hütter, Michael Schmidt, and Andreas Schwarte. The information workbench as a self-service

platform for developing linked data applications. *WWW 2012 Developer Track*, pages 18–20, 2012.

[HHU+11] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with swse: The semantic web search engine. *Web semantics: science, services and agents on the world wide web*, 9(4):365–401, 2011.

[HLS+08] Peter Haase, Holger Lewen, Rudi Studer, Duc Thanh Tran, Michael Erdmann, Mathieu d'Aquin, and Enrico Motta. The neon ontology engineering toolkit. *WWW*, 2008.

[HMK05] David Huynh, Stefano Mazzocchi, and David Karger. Piggy bank: Experience the semantic web inside your web browser. In *International Semantic Web Conference*, pages 413–430. Springer, 2005.

[HOAG10] AW Hill, J Otegui, AH Ariño, and RP Guralnick. Gbif position paper on future directions and recommendations for enhancing fitness-for-use across the gbif network, version 1.0. *Copenhagen: Global Biodiversity Information Facility*, 25:14, 2010.

[HP06] Jerry R Hobbs and Feng Pan. Time ontology in owl. *W3C working draft*, 27:133, 2006.

[HQD15] Wei Hu, Honglei Qiu, and Michel Dumontier. Link analysis of life science linked data. In *International Semantic Web Conference*, pages 446–462. Springer, 2015.

[HVOH06] Michiel Hildebrand, Jacco Van Ossenbruggen, and Lynda Hardman. /facet: A browser for heterogeneous semantic web repositories. In *International Semantic Web Conference*, pages 272–285. Springer, 2006.

[HZL08] Philipp Heim, Jürgen Ziegler, and Steffen Lohmann. gfacet: A browser for the web of data. In *Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08)*, volume 417, pages 49–58. Citeseer, 2008.

[IBM] Ibm watson. https://www.ibm.com/watson. [Online; accessed 20-May-2019].

[JHA+14] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman. Five stars of Linked Data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.

[KAC+14] 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.

[KD08] Georgi Kobilarov and Ian Dickinson. Humboldt: Exploring linked data. *context*, 6:7, 2008.

[KFS08] Jörg Koch, Thomas Franz, and Steffen Staab. Lena-browsing rdf data more complex than foaf. In *International Semantic Web Conference (Posters & Demos)*, 2008.

[KKS14] Yevgeny Kazakov, Markus Krötzsch, and František Simančík. The incredible elk. *Journal of automated reasoning*, 53(1):1–61, 2014.

[KSR+09] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media meets semantic web–how the bbc uses dbpedia and linked data to make connections. In *European semantic web conference*, pages 723–737. Springer, 2009.

[LKN+13] Yuan-Fang Li, Gavin Kennedy, Faith Ngoran, Philip Wu, and Jane Hunter. An ontology-centric architecture for extensible scientific data management systems. *Future Generation Computer Systems*, 29(2):641–653, 2013.

[LLC12] Adam Leadbetter, Roy Lowry, and DO Clements. The nerc vocabulary server: Version 2.0. In *Geophysical Research Abstracts*, volume 14, 2012.

[LOD19] The open linked data cloud. https://lod-cloud.net/, 2019. [Online; accessed 8-May-2019].

[Mar09] Marbles linked data engine. http://mes.github.io/marbles/, 2009. [Online; accessed 20-May-2019].

[MB09] Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. *W3C recommendation*, 18:W3C, 2009.

[MKMB+18] Frank E Muller-Karger, Patricia Miloslavich, Nicholas J Bax, Samantha Simmons, Mark J Costello, Isabel Sousa Pinto, Gabrielle Canonico, Woody Turner, Michael Gill, Enrique Montes, et al. Advancing marine biological observations and data requirements of the complementary essential ocean variables (eovs) and essential biodiversity variables (ebvs) frameworks. *Frontiers in Marine Science*, 5:211, 2018.

[MSG+16] João Luiz Rebelo Moreira, Tiago Prince Sales, John Guerson, Bernardo Ferreira Bastos Braga, Freddy Brasileiro, and Vinicius Sobral. Menthor editor: An ontology-driven conceptual modeling platform. In *Proceedings of the Joint Ontology Workshops 2016 Episode 2: The French Summer of Ontology co-located with the 9th International Conference on Formal Ontology in Information Systems (FOIS 2016), Annecy, France, July 6-9, 2016.*, 2016.

[Mus15] Mark A. Musen. The protÉgÉ project: A look back and a look forward. *AI Matters*, 1(4):4–12, June 2015.

[NF15] Tom Narock and Peter Fox. *The Semantic Web in Earth and space science. Current status and future directions*, volume 20. IOS Press, 2015.

[NSW+09] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173, 2009.

[ntr14] Rdf 1.1 n-triples: A line-based syntax for an rdf graph. https://www.w3.org/TR/n-triples/, 2014.

[OBI19] Ocean biogeographic information system 2.0. https://obis.org/, 2019. [Online; accessed 27-May-2019].

[OWL09] Web ontology language (owl). https://www.w3.org/OWL/, 2009. [Online; accessed 15-May-2019].

[OWL12] Owl 2 web ontology language. https://www.w3.org/TR/owl2-profiles/, 2012. [Online; accessed 16-May-2019].

[Pag19] Roderic DM Page. Ozymandias: A biodiversity knowledge graph. *PeerJ*, 7:e6739, 2019.

[PD03] Richard G Pearson and Terence P Dawson. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global ecology and biogeography*, 12(5):361–371, 2003.

[PH12] Matthew Perry and John Herring. Ogc geosparql-a geographic query language for rdf data. *OGC implementation standard*, 2012.

[PPMT18] Maria-Evangelia Papadaki, Panagiotis Papadakos, Michalis Mountantonakis, and Yannis Tzitzikas. An interactive 3d visualization for the lod cloud. In *EDBT/ICDT Workshops*, pages 100–103, 2018.

[PSHS11] Igor O Popov, MC Schraefel, Wendy Hall, and Nigel Shadbolt. Connecting the dots: a multi-pivot approach to data exploration. In *International semantic web conference*, pages 553–568. Springer, 2011.

[QL08] Bastian Quilitz and Ulf Leser. Querying distributed rdf data sources with sparql. In *European semantic web conference*, pages 524–538. Springer, 2008.

[RD11] K Döring M Robertson T Remsen D, Braak. Darwin Core Archive How-To Guide. 2011.

[RDF14] Rdf 1.1 n-triples: A line-based syntax for an rdf graph. https://www.w3.org/TR/rdf-syntax-grammar/, 2014. [Online; accessed 8-May-2019].

[RDG+14] 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.

[RJS11] O James Reichman, Matthew B Jones, and Mark P Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018):703–705, 2011.

[RVOH05] Lloyd Rutledge, Jacco Van Ossenbruggen, and Lynda Hardman. Making rdf presentable: integrated global and local semantic web browsing. In *Proceedings of the 14th international conference on World Wide Web*, pages 199–206. ACM, 2005.

[Sch09] Heiko Schuldt. Multi-tier architecture. In *Encyclopedia of database systems*, pages 1862–1865. Springer, 2009.

[SDC+14] Brian J Stucky, John Deck, Tom Conlin, Lukasz Ziemba, Nico Cellinese, and Robert Guralnick. The biscicol triplifier: bringing biodiversity data to the semantic web. *BMC bioinformatics*, 15(1):257, 2014.

[Sin13] Emily Singer. Biology's big problem: there's too much data to handle. *Quanta Magazine. Retrieved Jan*, 26:2014, 2013.

[SLHA12] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354, 2012.

[SPA08] Sparql query language for rdf. https://www.w3.org/TR/rdf-sparql-protocol/, 2008. [Online; accessed 8-May-2019].

[SPA13] Sparql 1.1 query language. https://www.w3.org/TR/sparql11-query/, 2013. [Online; accessed 15-May-2019].

[SPG+07] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 5(2):51–53, 2007.

[SSF+18] Viktor Senderov, Kiril Simov, Nico Franz, Pavel Stoev, Terry Catapano, Donat Agosti, Guido Sautter, Robert A Morris, and Lyubomir Penev. Openbiodiv-o: ontology of the openbiodiv knowledge management system. *Journal of biomedical semantics*, 9(1):5, 2018.

[STH10] Thomas Steiner, Raphaël Troncy, and Michael Hausenblas. How google is using linked data today and vision for tomorrow. 2010.

[STH+19] Derrick Snowden, Vardis M. Tsontos, Nils Olav Handegard, Marcos Zarate, Kevin O' Brien, Kenneth S. Casey, Neville Smith, Helge Sagen, Kathleen Bailey, Mirtha N. Lewis, and Sean C. Arms. Data interoperability between elements of the global ocean observing system. *Frontiers in Marine Science*, 6:442, 2019.

[TDO07] Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice. com: Weaving the open linked data. In *The Semantic Web*, pages 552–565. Springer, 2007.

[Tim09] Putting government data online. https://www.w3.org/DesignIssues/GovData.html, 2009. [Online; accessed 17-May-2019].

[TPR10] Edward Thomas, Jeff Z Pan, and Yuan Ren. Trowl: Tractable owl 2 reasoning infrastructure. In *Extended Semantic Web Conference*, pages 431–435. Springer, 2010.

[tur14] Rdf 1.1 turtle: Terse rdf triple language. https://www.w3.org/TR/turtle/, 2014. [Online; accessed 8-May-2019].

[VDW13] Ruben Verborgh and Max De Wilde. *Using OpenRefine*. Packt Publishing Ltd, 2013.

[VG06] Denny Vrandečić and Aldo Gangemi. Unit tests for ontologies. In *OTM Confederated International Conferences On the Move to Meaningful Internet Systems*, pages 1012–1020. Springer, 2006.

[VTVBCGP11] Boris Villazón-Terrazas, Luis M Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez. Methodological guidelines for publishing government linked data. In *Linking government data*, pages 27–49. Springer, 2011.

[W3c01] Uris, urls, and urns: Clarifications and recommendations 1.0. https://www.w3.org/TR/uri-clarification/, 2001. [Online; accessed 27-May-2019].

[W3C04] W3C. Resource description framework. https://www.w3.org/TR/rdf-concepts/, 2004. [Online; accessed 8-May-2019].

[WBG+12] 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.

[WDA+16] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[WDG+14] Ramona L Walls, John Deck, Robert Guralnick, Steve Baskauf, Reed Beaman, Stanley Blum, Shawn Bowers, Pier Luigi Buttigieg, Neil Davies, Dag Endresen, et al. Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. *PLoS One*, 9(3):e89606, 2014.

[WT13] Jesse Weaver and Paul Tarjan. Facebook linked data via the graph api. *Semantic Web*, 4(3):245–250, 2013.

[XML08] Extensible markup language (xml) 1.0 (fifth edition). https://www.w3.org/TR/xml/, 2008. [Online; accessed 8-May-2019].

[YBS+07] Chris Yesson, Peter W Brewer, Tim Sutton, Neil Caithness, Jaspreet S Pahwa, Mikhaila Burgess, W Alec Gray,

Richard J White, Andrew C Jones, Frank A Bisby, et al. How global is the global biodiversity information facility? *PLoS One*, 2(11):e1124, 2007.

[ZBF] Marcos Zárate, Germán Braun, and Pablo Fillottrani. Adding biodiversity datasets from argentinian patagonia to the web of data.

[ZBF17] Marcos Zárate, Agustina BUCCELLA, and Pablo FILLOTTRANI. Bioonto: Towards an integration of biological and biogeographic data. In *Ontologies and Data in Life Sciences (ODLS 2017), in conjunction with the Joint Ontology Workshops 2017*, 2017.

[ZBF+19] Marcos Zárate, Gemán Braun, Pablo Rubén Fillottrani, Claudio Delrieux, and Mirtha Lewis. Bige-onto: An ontology-based system for managing biodiversity and biogeography data. accepted paper, 2019.

[ZL16] M D Zárate and M N Lewis. Estimate of the Anesthesia Stage in Southern Elephant Seals using WEKA Data Mining Tool. *International Journal of Applied Information Systems*, 11(4), 2016.

[ZLFD18] Marcos Zárate, Mirtha Lewis, Pablo Rubén Fillottrani, and Claudio Delrieux. Improving the quality of biodiversity data through semantic web standards. In Jitendra Gaikwad, Birgitta König-Ries, Friedrich Recknagel, et al., editors, *ICEI 2018: 10th International Conference on Ecological Informatics-Translating Ecological Data into Knowledge and Decisions in a Rapidly Changing World*, 2018.

[ZRB+19] Marcos Zárate, Pablo Rosales, Germán Braun, Mirtha Lewis, Pablo Rubén Fillottrani, and Claudio Delrieux. Oceangraph: Some initial steps toward a oceanographic knowledge graph. In Boris Villazón-Terrazas and Yusniel Hidalgo-

Delgado, editors, *Knowledge Graphs and Semantic Web*, pages 33–40, Cham, 2019. Springer International Publishing.

[ZRF+18] Marcos Zárate, Pablo Rosales, Pablo Fillottrani, Claudio Delrieux, and Mirtha Lewis. Oceanographic data management: Towards the publishing of pampa azul oceanographic campaigns as linked data. 2018.