Herramientas Informáticas para el Estudio de la Biodiversidad utilizando Datos Abiertos Enlazados

Gustavo Samec[1,2,3], María Emilia Diez[1,2,5], Marcos Zárate[1,2,4], Carlos Buckle[1,2], Joaquín Lima[1,2], Rodrigo Jaramillo[1,2,3], Alejandro Sánchez[1,2,6], Renato Mazzanti[1,2,3]

[1] Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Puerto Madryn.
[2] LINVI, Laboratorio de Investigación en Informática, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB)
[3] Unidad de Gestión de la Información, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (CCT CONICET-CENPAT)
[4] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (CESIMAR) (CCT CONICET-CENPAT)
[5] Laboratorio de Parasitología (LAPA), Instituto de Biología de Organismos Marinos, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (IBIOMAR) (CCT CONICET-CENPAT)
[6] Departamento de Informática, Universidad Nacional de San Luis (UNSL), San Luis.
{gsamec, emiliadiez, zarate, rjaramillo, renato}@cenpat-conicet.gob.ar, cbuckle@unpata.edu.ar

RESUMEN

En la actualidad existen grandes bases de datos globales de biodiversidad con contenidos abiertos a la comunidad científica, la mayoría proveen APIs Web para realizar consultas y recuperar información. A pesar de éstas facilidades frecuentemente no son interoperables, por lo general representan un modelo de datos propietario y carecen de vocabularios de descripciones semánticas formales esenciales para garantizar la integración de los datos. El presente trabajo tiene como objetivo hacer accesible y abiertos los datos, a la comunidad científica, de la base de datos Southwest Atlantic Benthic Invertebrates[1] (SWATL) que registra datos de de invertebrados bentónicos de la región y publicaciones taxonómicas, por medio de Datos Abiertos Enlazados[2] (LOD) y hacerla interoperable con bases de datos de referencia global desarrollando micro-servicios SPARQL[3] como envoltura (wrapper) de las APIs Web que las mismas proveen.

Palabras clave: Web Semántica, Datos Abiertos Enlazados, SPARQL, Biodiversidad.

CONTEXTO

Este trabajo se encuadra dentro del proyecto "Aplicaciones Informáticas para el Estudio de Biodiversidad de Poliquetos Espiónidos en los Golfos Nordpatagónicos", elaborado en el Laboratorio de Investigación en Informática (LINVI) de la Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB) e integrado por docentes investigadores de la Facultad de Ingeniería Sede Puerto Madryn, con participación de estudiantes y graduados de las carreras de dicha Sede. El proyecto fue avalado por el Consejo Directivo de la Facultad de Ingeniería de la UNPSJB y es financiado por la Secretaría de Ciencia y Técnica de la Universidad para llevarlo a cabo durante 2019-2020.

1. INTRODUCCION

La gran diversidad de datos generados por distintas disciplinas (taxonómicos, genéticos, meteorológicos, etc.) y la posibilidad de extraerlos e integrarlos para realizar investigaciones resulta de sumo

[1] http://sistema.cenpat-conicet.gob.ar:8081/rb/
[2] https://www.w3.org/egov/wiki/Linked_Open_Data
[3] https://www.w3.org/TR/rdf-sparql-query/

interés a la comunidad científica [1]. La Web Semántica [2] proporciona soluciones a éstas necesidades a través de Datos Enlazados (Linked Data) [3] donde los objetos de datos se identifican de manera única y las relaciones entre ellos se definen explícitamente.

En nuestro caso, contamos con la base de datos SWATL sobre la cual se requieren desarrollar herramientas informáticas para que sus datos sean accesibles y abiertos a la comunidad científica y por otro lado interoperable con otras bases de datos de referencia global permitiendo hacer consultas y extraer información de ellas.

Si bien es cierto que existen grandes bases de datos con contenidos similares como por ejemplo: Global Biodiversity Information Facility[4] (GBIF), World Register of Marine Species[5] (WoRMS), Species 2000[6] (SP2000), etc., que gestionan listas taxonómicas de especies a partir de la validación de especialistas taxónomos, la importancia de SWATL radica en que es una base de datos regional focalizada a un grupo determinado de especies y publicaciones taxonómicas, alimentada y mantenida por especialistas locales, que en numerosos casos cuenta con más información y está más actualizada que las grandes bases de datos de referencia a nivel global, siendo su contenido de gran interés por quienes desarrollan investigaciones que tienen relación con los datos que almacena la misma.

Para el acceso abierto de los datos de SWATL se propone utilizar LOD, para ello se requiere exportar el contenido de la base de datos relacional a tripletas RDF[7]. La utilización de LOD mejora la capacidad de localización, accesibilidad, interoperabilidad y reutilización de los mismos (principios FAIR - Findable, Accessible, Interoperable, and Reusable) [4], y permite la vinculación con otros conjuntos de datos abiertos enlazados.

Esta vinculación de datos, con el agregado de una mayor integración semántica [5], facilita a los científicos descubrir y utilizar esos datos de manera consistente en sus investigaciones.

A la hora de hacer interoperable SWATL con otras bases de datos encontramos la limitación que, en su gran mayoría, no proveen un endpoint SPARQL. Por lo general los portales de acceso a las bases de datos proveen APIs Web para hacer distintos tipos de consultas y extraer registros de los millones de datos que almacenan. En este sentido realizar consultas interoperables con SWATL demandaría realizar un código *ad-hot* para cada tipo de proveedor.

En este caso la propuesta es desarrollar micro-servicios SPARQL [6,7] como envoltura de las APIs Web que los portales de las bases proveen y traducir su respuesta en tripletas RDF lo cual va a permitir que SWATL interopere con las mismas.

2. LÍNEAS DE INVESTIGACION Y DESARROLLO

Dentro de los trabajos realizados, se rediseño el modelo de la base de datos SWATL, en el mismo se incluyeron datos que facilitan la integración principalmente con WoRMS y bases relacionadas que utilizan la Plataforma Aphia [8] y se reorganizaron sus tablas y relaciones para simplificar las consultas y obtener información con un contenido similar a las ofrecidas por las base de datos de referencia global. Por otro lado se realizó una refactorización de la aplicación que accede a la base de datos para su carga y consulta y los cambios necesarios para adaptarla al nuevo modelo de datos. El desarrollo original corresponde a una aplicación Web desarrollada por capas, todos sus módulos están escritos en Java y utiliza el estándar JavaServer Faces (JSF) a nivel de capa de presentación. Las capas de modelo, acceso a datos y servicios se adaptaron a los cambios requeridos por las nuevas bibliotecas utilizadas además de los cambios correspondientes al nuevo modelo de datos. La capa lógica prácticamente no requirió cambios más allá de la adaptación al nuevo

[4] https://www.gbif.org/
[5] http://www.marinespecies.org/
[6] https://sp2000.org/
[7] https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

modelo de datos. La capa de presentación demandó un gran esfuerzo, si bien se siguió utilizando el mismo estándar JSF, se cambió la utilización de componentes de ICEFaces[8], por Primefaces[9]. La utilización de componentes más avanzados provisto por Primefaces motivó cambios significativos en esta capa. También se cambió la tecnología utilizada en seguridad, la versión original utilizaba Realm de Tomcat y la nueva versión utiliza SpringSecurity independizándola del servidor de aplicaciones. Las razones que motivaron la refactorización de la aplicación original responde a adaptarlas a las nuevas tecnologías existentes permitiendo que la aplicación actual y los futuros desarrollos sobre la misma capitalicen las ventajas que éstas proveen.

Por otro lado se han realizado experiencias utilizando las APIs que provee Apache Jena[10] y Graphdb[11] para exportar el contenido de la base de datos relacional a LOD y permitir consultarla en un endpoint SPARQL. Las bases de datos taxonómicas y de registros históricos como SWATL tiene una baja tasa de actualización por lo que tener una copia de la base de datos relacional (BDR) en un *triplestore* y actualizarla periódicamente no afectaría a la calidad del servicio, de todas maneras no se descarta evaluar la performance y utilizar servidores que acceden a la BDR y convierten los datos en LOD sin necesidad de almacenarlos *(on-fly)*.

Para interoperar con otras bases de datos (que no posean endpoint SPARQL) la propuesta es desarrollar micro-servicios SPARQL. Para ello se requiere crear un micro-servicio SPARQL por cada API Web que la base provee. Dada una consulta, el cliente SPARQL envía la misma al micro-servicio SPARQL, el cual se encarga de traducirla al formato adecuado y enviar la solicitud a la API Web, cuando la misma responde, el micro-servicio SPARQL se encarga de traducir el resultado en tripletas RDF y enviar las mismas al cliente SPARQL.

El starndard Darwin Core [9] es ampliamente utilizado en bases de datos de biodiversidad (incluido SWATL), el hecho de poseer un vocabulario en común facilita la interoperabilidad.

3. RESULTADOS ESPERADOS

Poner a disposición de la comunidad científica los datos de SWATL y facilitar su interoperabilidad con bases de datos de referencia global.

Aplicar la experiencia adquirida y los productos obtenidos a otros campos donde se requiera interoperabilidad.

Contribuir en la formación de recursos humanos en nuevas tecnologías.

Consolidar un grupo de investigación interdisciplinario en informática para la biodiversidad dentro del grupo LINVI de la UNPSJB.

4. FORMACION DE RECURSOS HUMANOS

En este proyecto participan integrantes de formación docente y académica en las áreas de Ingeniería de Software, Bases de Datos, Inteligencia Artificial y Biología. Cinco de los docentes son del Departamento de Informática y uno del Departamento de Matemáticas de la Facultad de Ingeniería de la UNPSJB Sede Puerto Madryn. La Doctora en biología María Emilia Diez es investigadora del CONICET y es especialista en taxonomía y ecología de poliquetos, uno de los autores están realizando la carrera de doctorado, otro iniciando su post-doctorado con beca del CONICET y otros dos se encuentran desarrollando carreras de especialización y maestrías. También forman parte del grupo de investigación un graduado de la carrera de Licenciatura en Informática y tres alumnos del ciclo superior. Uno de ellos desarrollando su tesina.

[8] http://www.icesoft.org/

[9] https://www.primefaces.org/

[10] https://jena.apache.org/

[11] http://graphdb.ontotext.com/documentation/free/quick-start-guide.html

5. BIBLIOGRAFIA

1. Richard K Lomotey and Ralph Deters. Terms extraction from unstructured data silos. In System of Systems Engineering (SoSE), 2013 8th International Conference on, pages 19-24. IEEE, 2013.

2. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. Scientific American, 284(5):28-37, 2001.

3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In Semantic services, interoperability and web applications: emerging concepts, pages 205-227. IGI Global, 2011.

4. 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.

5. Paolo Ceravolo, Antonia Azzini, Marco Angelini, Tiziana Catarci, Philippe Cudr_e-Mauroux, Ernesto Damiani, Alexandra Mazak, Maurice Van Keulen, Mustafa Jarrar, Giuseppe Santucci, et al. Big data semantics. Journal on Data Semantics, 7(2):65-85, 2018.

6. Michel F, Faron-Zucker C, Tercerie S, Ettorre A, Olivier G. Assisting Biologists in Editing TaxonomicInformation by Confronting Multiple Data Sources using Linked Data Standards. Biodiversity Information Science and Standards 3, 2019. https://doi.org/10.3897/biss.3.37421

7. Michel F, Faron-Zucker C, Gargominy O, Gandon F. Integration of Web APIs and Linked Data Using SPARQL Micro-Services - Application to Biodiversity Use Cases, 2018. https://doi.org/10.3390/info9120310

8. Nozères, C., Vandepitte, L., Appeltans, W., Kennedy, M. Best Practice Guidelines in the Development and Maintenance of Regional Marine Species Checklists, version 1.0, released on August 2012. Copenhagen: Global Biodiversity Information Facility, 32 pp. 2012 http://www.gbif.org/orc/?doc_id=4712

9. Baskauf S, Wieczorek J, Deck J, Webb C, Morris PJ, Schildhauer M. Darwin Core RDF Guide. Biodiversity Information Standards (TDWG). 2015. http://rs.tdwg.org/dwc/terms/guides/rdf/