Transformando Datos de Biodiversidad en Linked Data

Marcos Zarate[12], German Braun[3], Samuel Almonacid[2]

[1] Laboratorio de Investigación en Informática (LINVI), Facultad de Ingeniería
Universidad Nacional de la Patagonia San Juan Bosco.
[2] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico Consejo Nacional de Investigaciones Científicas y Técnicas (CESIMAR-CENPAT-CONICET)
[3] Grupo de Investigación en Lenguajes e Inteligencia Artificial Departamento de Teoría de la Computación
Facultad de Informática, Universidad Nacional del Comahue
{zarate,almonacid}@cenpat-conicet.gob.ar, german.braun@fi.uncoma.edu.ar

Resumen

La biodiversidad es esencial para la vida en la tierra y motiva muchos esfuerzos en la recopilación de datos sobre especies, que son utilizados por investigadores para el estudio de los seres vivos. Sin embargo, dados que estos datos se extraen desde diferentes lugares geográficos y se almacenan en distintos formatos, su recuperación, combinación e integración es aún un problema abierto.

El objetivo general de este trabajo de investigación es desarrollar una arquitectura para convertir y publicar datos de biodiversidad utilizando tecnologías de la Web Semántica, en particular los principios establecidos por la iniciativa Linked Open Data (LOD) para compartir y relacionar información. Esta línea de investigación se desarrolla en forma colaborativa entre docentes-investigadores de la Universidad Nacional del Comahue y de la Universidad Nacional de la Patagonia San Juan Bosco, en el marco de proyectos de investigación financiados por las universidades antes mencionadas.

Palabras clave: Linked Data, Web Semántica, Biodiversidad, RDF

Contexto

El LINVI es el Laboratorio de Investigación en Informática dependiente de la Facultad de Ingeniería de la Universidad Nacional de la Patagonia San Juan Bosco. La especialidad del laboratorio es la investigación, desarrollo, servicios de vinculación y formación de recursos humanos en Informática, en particular esta línea de investigación se incluye dentro del proyecto *Clasificación de Información en BigData mediante la utilización de Técnicas de Inteligencia Artificial y Análisis de Redes Sociales*.

Este trabajo también está parcialmente financiado por la Universidad Nacional del Comahue, en el marco del proyecto de investigación *Agentes Inteligentes y Web Semántica*, cuya duración es de cuatro años, y por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), en el contexto de dos becas internas doctorales de 5 anos de duración, finalizando en abril de 2019 y 2020.

Por otra parte, el CESIMAR unidad ejecutora dependiente del CONICET desarrolla líneas de investigación y forma recursos humanos orientados a comprender el funcionamiento de los ecosistemas marinos, y brindar soporte científico y tecnológico para el manejo y la conservación de los recursos del mar. Estas entidades cuentan con Convenios de Colaboración mutua y han realizado

trabajos conjuntos en temáticas relacionadas con los sistemas de tiempo real, inteligencia artificial, redes y procesamiento de imágenes.

1. Introducción

Las colecciones biológicas contienen información heredada irreemplazable sobre nuestra biosfera, que es esencial para comprender como la biodiversidad está cambiando en una era de impactos humanos sin precedentes. Dichos análisis solo son prácticos si se digitalizan, integran y hacen disponibles en línea los datos de las colecciones biológicas de todo el mundo. Estas tareas son un foco importante del campo de la informática y, aunque presentan muchos desafíos, también prometen ofrecer beneficios significativos para la ciencia que estudia la biodiversidad y sus disciplinas asociadas. En los últimos años, la comunidad informática de la biodiversidad ha hecho grandes progresos hacia el logro de este objetivo mediante la creación de vocabularios comunes compartidos, como Darwin Core (DwC) [1] y mecanismos de publicación como el Integrated Publishing Toolkit (IPT) [2]. Gracias a estas y otras iniciativas nacionales e internacionales, ahora existen cientos de millones de registros de biodiversidad de todo el mundo.

Aunque se han conseguido logros sustanciales en llegar a un consenso sobre los estándares de datos, formatos de metadatos y mecanismos de intercambio por parte de las organizaciones de normalización como el Open Geospatial Consortium, Inc. (OGC[1]), la Organizacion Internacional de Normalización (ISO[2]) y el World Wide Web Consortium (W3C[3]), todavía hay una serie de vacíos que impiden la plena interoperabilidad entre los sistemas de información de biodiversidad. Los científicos son desafiados por el volumen y la heterogeneidad de los tipos de datos y formatos [3], y la dificultad de descubrir, acceder e integrar conjuntos de datos desde múltiples fuentes [4,5].

La Web Semántica [6] y sus tecnologías asociadas proporcionan una solución natural a estos problemas al permitir una red de datos y conocimientos vinculados donde todos los objetos de datos son identificados de manera única y las relaciones entre ellos están explícitamente definidas [6,7]. Por consiguiente, cada vez son más reconocidas las ventajas de las tecnologías asociadas a la Web Semántica y a LOD, no solo en la investigación sobre la diversidad biológica y sus disciplinas conexas (por ejemplo, [8,9,10,11]), sino también en las ciencias de la vida [12,13,14].

En la actualidad uno de los portales de biodiversidad más utilizados por la comunidad científica internacional pertenece a la organización intergubernamental Global Biodiversity information Facility (GBIF[4]), cuyo objetivo es estructurar nodos nacionales en una red accesible vía Web de manera libre y gratuita. GBIF nace en 2001 y comprende en la actualidad 53 países y 43 organizaciones internacionales.

El nodo Argentino de GBIF es el Sistema Nacional de Datos Biológicos (SNDB[5]), una iniciativa del Ministerio de

[1] http://www.opengeospatial.org

[2] http://www.iso.org/iso/home.html

[3] https://www.w3.org/

[4] http://www.gbif.org/

[5] http://datos.sndb.mincyt.gob.ar/

Ciencia, Tecnología e Innovación Productiva, conjuntamente con el Consejo Interinstitucional de Ciencia y Tecnología (CICyT) y enmarcada dentro del Programa de Grandes Instrumentos y Bases de Datos. Sin embargo, aunque GBIF soporta interoperabilidad a través de su estándar para datos de biodiversidad Darwin Core, en la actualidad, estos datos no están publicados siguiendo los criterios establecidos por la iniciativa LOD, lo que representa una desventaja a la hora de compartir e integrar información. El objetivo de esta línea de investigación es estudiar e implementar modelos y arquitecturas que permitan la publicación de datos de biodiversidad en la Web de Datos. En el ámbito de este trabajo proponemos analizar diferentes tecnologías y formalismos que se requieren para poder ofrecer una fuente de datos enlazada y abierta, con información de biodiversidad, como así también enfoques para su integración y herramientas para consumir estos datos que se beneficien de su forma estandarizada de representación y de la posibilidad de enlazar nuevas fuentes de datos en tiempo de ejecución. En particular, nos centraremos en la preparación y conversión de los datos a RDF [15], la definición de las URIs y el interenlazado de nuestras fuentes actuales. Como trabajo futuro, planeamos desarrollar un prototipo de aplicación Web especıfica para visualizar estos datos y así, aumentar la visibilidad fomentando la colaboración entre grupos interdisciplinarios.

El presente trabajo se estructura de la siguiente forma. En la sección 2 presentamos los objetivos de los proyectos de investigación en los que se enmarca este trabajo y describimos la línea de investigación, el problema que se estudia y los objetivos. En la sección 3 indicamos algunos resultados obtenidos y trabajos futuros. Finalmente, comentamos aspectos referentes a la formación de recursos humanos en esta temática

2. Líneas de Investigación, Desarrollo e Innovación

El proyecto de investigación *Clasificación de Información en BigData mediante la utilización de Técnicas de Inteligencia Artificial y Análisis de Redes Sociales* tiene como objetivo evaluar técnicas existentes e implementar desarrollos experimentales que permitan clasificar, ordenar, jerarquizar y analizar información sobre grandes volúmenes utilizando tecnologías semánticas.

Por otro lado, el proyecto de investigación *Agentes Inteligentes y Web Semántica* tiene varios objetivos generales. Uno de ellos es el de desarrollar conocimiento especializado en el área de Interoperabilidad Semántica de la Información. En este sentido, se estudian, entre otras, el desarrollo de agentes de información cuyo ambiente de trabajo es la Web, además de técnicas de representación de conocimiento, razonamiento automático y modelado ontológico.

Estas líneas de investigación confluyen en el estudio de formalismos y tecnologías para cubrir las necesidades emergentes de compartir, actualizar e integrar el conocimiento de sistemas computacionales pre-existentes. Particularmente, hemos escogido experimentar en la publicación de datos de biodiversidad para generar datos abiertos y enlazados con otras fuentes disponibles en la Web. Para ello, se utilizaran conjuntos de datos científicos existentes de investigadores colaboradores del Centro Nacional

Patagónico (CENPAT-CONICET), con la finalidad de contar con expertos que nos permitan realizar y estructurar los mecanismos de clasificación y evaluar los resultados de los desarrollos experimentales en estos conjuntos de datos.

3. Resultados Obtenidos y Trabajo Futuro

Inicialmente, se realizó un relevamiento de las tecnologías disponibles para nuestra arquitectura, y se puntualizó sobre las plataformas D2RQ [16], Jena [17], OpenRefine [18] y GraphDB [19], entre otras. De este análisis, se determinó que las plataformas más convenientes para un primer prototipo son OpenRefine y GraphDB. El primero soporta extensiones para la creación de triplas RDF[6] a partir de una gran variedad de formatos de entrada, tales como CSV, hojas de cálculo, JSON y el mismo formato RDF. Además, permite explorar y depurar los conjuntos de datos, aplicar transformaciones y definir vocabularios asociados a los diferentes campos, de una manera amigable. Por otro lado, GraphDB es un repositorio semántico que permite almacenar las tripletas generadas por OpenRefine y también trabajar con motores de inferencia y de consultas SPARQL [20] sobre estos datos estructurados.

La Fig. 1 muestra la arquitectura inicial para la conversión de los datos. Dicha arquitectura acepta datos de biodiversidad en formato tabular estándar, incluyendo DwC-A. El proceso de depuración y exploración de datos involucra la definición de URIs y los diferentes vocabularios basados, principalmente, en el estándar Darwin Core[7] y aquellos para la descripción de datos personales, filiatorios y de recursos, tales como FOAF[8] y Dublin Core[9], respectivamente. Este proceso, inicialmente manual, genera una salida en formato JSON, para procesar los siguientes repositorios de manera automática.

Actualmente, se está trabajando en la vinculación de nuestros repositorios, como un proceso de interenlazado interno, de depuración y exploración, y con otros repositorios RDF externos. En todos los casos, se está experimentando con LIMES[10] y SILK[11], las cuales son herramientas automáticas para la detección de enlaces en la Web de Datos. Finalmente, está previsto diseñar un prototipo para la visualización y consulta de los datos generados y enlazados, además de su ontología relacionada. Dicho prototipo estará basado en la herramienta para modelado ontológico crowd [21].

4. Formación de Recursos Humanos

Dos de los autores de este trabajo están inscriptos en el Doctorado en Ciencias de la Computación en la Universidad Nacional del Sur, mientras que uno de ellos se encuentra inscripto en el Doctorado en Ingeniería Mención en Procesamiento de Señales e Imágenes en

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

[7] http://rs.tdwg.org/dwc/rdf/dwcterms.rdf

[8] http://xmlns.com/foaf/spec/

[9] http://dublincore.org/documents/dcmi-terms/

[10] http://aksw.org/Projects/LIMES.html

[11] http://silkframework.org/

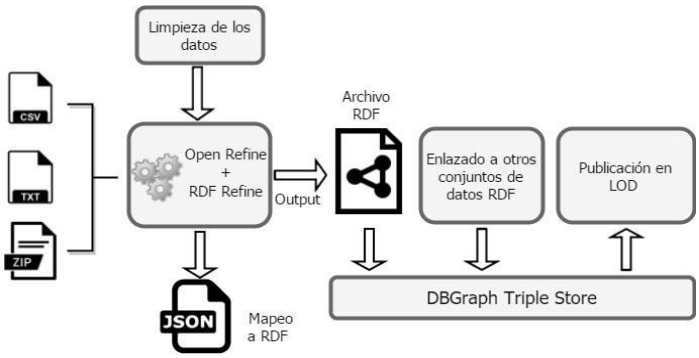

Fig. 1. Arquitectura inicial para la transformación de datos de biodiversidad a RDF.

la Universidad Tecnológica Nacional Facultad Regional Buenos Aires. Los mismos cuentan con beca interna doctoral del CONICET. El presente trabajo se enmarca dentro de la investigación realizada para el desarrollo de las tesis doctorales, en donde se investigan las diversas formas de integrar conjuntos de datos de biodiversidad de diversos repositorios digitales usando tecnologías de la Web Semántica. El tema que se presenta viene profundizándose mediante el estudio continuo y con la presentación de diferentes trabajos en reuniones científicas en donde se muestran los avances realizados y los posibles resultados que se esperan de la investigación.

Referencias

[1] John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Doring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin core: an evolving community-developed biodiversity data standard. PloS one, 7(1):e29715, 2012. Author. (year, month). Title. Presented at Conference title. [Type of Medium]. Available: site/path/file

[2] Tim Robertson, Markus Doring, Robert Guralnick, David Bloom, John Wieczorek, Kyle Braak, Javier Otegui, Laura Russell, and Peter Desmet. The gbif integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PLoS One, 9(8):e102623, 2014.

[3] Anne E Thessen and David J Patterson. Data issues in the life sciences. ZooKeys, 150(150):15–51, 2011.

[4] 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.

[5] Tanu Malik and Ian Foster. Addressing data access needs of the long-tail distribution of geoscientists. In Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International, pages 5348–5351. IEEE, 2012.

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

[7] 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.

[8] Roderic DM Page. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in bioinformatics, 9(5):345–354, 2008.

[9] Cynthia S Parr, Robert Guralnick, Nico Cellinese, and Roderic DM Page. Evolutionary informatics: unifying knowledge about the diversity of life. Trends in ecology & evolution, 27(2):94–103, 2012.

[10] Joshua S Madin, Shawn Bowers, Mark P Schildhauer, and Matthew B Jones. Advancing ecological research with ontologies. Trends in ecology & evolution, 23(3):159–168, 2008.

[11] Andrew R Deans, Matthew J Yoder, and James P Balhoff.Time to change how we describe biodiversity. Trends in ecology & evolution, 27(2):78–84, 2012.

[12] Robert Stevens, Carole A Goble, and Sean Bechhofer. Ontology-based knowledge representation for bioinformatics. Briefings in bioinformatics, 1(4):398– 414, 2000.

[13] Judith A Blake and Carol J Bult. Beyond the data deluge: data integration and bio-ontologies. Journal of biomedical informatics, 39(3):314–320, 2006.

[14] Huajun Chen, Tong Yu, and Jake Y Chen. Semantic web meets integrative biology: a survey. Briefings in bioinformatics, 14(1):109–125, 2013.

[15] Ora Lassila and Ralph R Swick. Resource description framework (rdf) model and syntax specification. 1999.

[16] Christian Bizer and Andy Seaborne. Treating non-rdf databases as virtual rdf graphs.

[17] Brian McBride. Jena: A semantic web toolkit. IEEE Internet computing, 6(6):55–59, 2002.

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

[19] Ralf Hartmut Guting. Graphdb: Modeling and querying graphs in databases. In VLDB, volume 94, pages 12–15, 1994

[20] Eric Prud, Andy Seaborne, et al. Sparql query language for rdf. 2006.
Christian Gimenez, German Braun, Laura Cecchi, and Laura Fillottrani. crowd: A Tool for Conceptual Modelling assisted by Automated Reasoning - Preliminary Report. In Proc. of the 2nd Simposio Argentino de Ontologías y sus Aplicaciones (SAOA) co- located at Jornadas Argentinas de Informatica (JAIIO) - to appear , 2016.