Extracción y Explotación de Conocimiento para la Gestión en línea de datos en Ciencias del Mar

Marcos Zárate[1,3], Carlos Buckle[1], Renato Mazzanti[1,2], Mirtha Lewis[3,4] Gustavo Nuñez[1] , Dario Ceballos[3]

[1] Laboratorio de Investigación en Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (LINVI-FI-UNPSJB), Puerto Madryn, Argentina,+54 280-4883585 Int. 117.

[2] Unidad de Gestión de la Información, Centro Nacional Patagónico,
Consejo Nacional de Investigaciones Científicas y Técnicas (UGI-CENPAT-CONICET), Puerto Madryn, Argentina

[3] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico,
Consejo Nacional de Investigaciones Científicas y Técnicas (CESIMAR-CENPAT-CONICET), Puerto Madryn, Argentina.

[4] Centro de Investigaciones y Transferencia Golfo San Jorge, (CIT-GSJ-CONICET), Comodoro Rivadavia, Argentina.

zarate@cenpat-conicet.gob.ar, cbuckle@unpata.edu.ar, renato@cenpat-conicet.gob.ar, mirtha@cenpat-conicert.gob.ar, guscostaf@gmail.com, disenodc@gmail.com

Resumen

El objetivo general de esta investigación es estudiar y desarrollar grafos de conocimiento para la gestión integrada y visualización de datos en línea de ciencias del mar a través de tecnologías Big Data y datos provenientes de campañas oceanográficas, de repositorios de Biodiversidad marina y de datos ambientales. En una primera etapa será acotado al golfo San Jorge (GSJ) dado que se cuenta con datos provenientes de las campañas realizadas por el grupo de trabajo GSJ, 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. El impacto esperado de esta investigación será el de permitir una gestión confiable de los datos y una explotación adecuada en línea, para garantizar la preservación y el acceso a nuestros activos nacionales de investigación multidisciplinar. El proyecto se realiza en conjunto entre LINVI-FI-UNPSJB y CESIMAR-CENPAT-CONICET. En el proyecto participan docentes investigadores de la carrera de Licenciatura en Informática, investigadores, un becario post-doctoral y un becario CONICET y un estudiante avanzado aspirante a beca EVC-CIN.

Palabras clave: Grafos de Conocimiento, Datos Abiertos Enlazados, Datos Oceanográficos, Grandes Volúmenes de Datos.

Contexto

Esta propuesta avanzará sobre resultados y líneas de investigación del proyecto precedente *UNPSJB-PI 1562 - Plataforma de Datos Abiertos Enlazados para la Gestión y Visualización de Datos Primarios de Ciencias del Mar*, en el cual se identificaron ventajas en los Datos Abiertos Enlazados (Linked Open Data) [1] como camino estándar hacia la integración de datos abiertos [2] heterogéneos de diferentes dominios, bajo vocabularios comunes y con posibilidades de razonamiento por partes de agentes de software. Este trabajo se focalizará específicamente en las estrategias y técnicas de visualización de datos científicos relacionados con las ciencias del mar para poder ser explotados de manera adecuada por los investigadores.

Estos proyectos tienen una concepción interdisciplinaria y por ello el grupo de trabajo incluye: investigadores de las ciencias de la

computación de diferentes áreas e investigadores de las ciencias biológicas, expertos en el dominio de aplicación.

Este proyecto está avalado por el Consejo Directivo de la Facultad de Ingeniería de la UNPSJB y será financiado por la Secretaría de Ciencia y Técnica de la UNPSJB para llevar a cabo durante el periodo 2022-2023.

1. Introducción

En el contexto de la iniciativa Pampa Azul [3] relanzada en julio de 2020, la gestión y modelado de datos en línea es considerada una disciplina fundamental para abordar la complejidad y el alcance de las temáticas que exigen una aproximación interdisciplinaria y una proyección amplia en el uso de la información. Por lo tanto, es necesario desarrollar sistemas capaces de gestionar su integración y su comunicación, tanto para un aprovechamiento integral y secundario de los grupos e instituciones participantes como para usuarios externos que requieran de información.

La relevancia de la propuesta se identifica en diferentes niveles: **(1)** a nivel tecnológico, la creación de una plataforma que permita gestionar datos de ciencias del mar, permitirá resolver los problemas de integración de datos heterogéneos bajo un mismo vocabulario y disponerlos para ser explotados mediante búsquedas semánticas y consultas basadas en lenguaje natural. Adicionalmente, el desarrollo de visualizaciones de datos relacionados a las ciencias del mar (Data Visualization) [3] facilitará su interpretación y comparación. **(2)** a nivel de política científica para la región y el país, se logrará un prototipo de servicio estandarizado para repositorios de datos marinos del Atlántico Sur Occidental, con información unificada provenientes de diferentes plataformas de muestreo (buques oceanográficos, ROVs, animales instrumentados, etc.), que podrá ser reusada y explotada para la generación de nuevos conocimientos. **(3)** a nivel socioeconómico, las posibilidades de reuso de datos permitirán reducir los altos costos de las campañas de investigación en el mar y aportará al uso democrático y eficiente de datos científicos. Finalmente **(4)** a nivel de formación de recursos humanos del proyecto participa un becario doctoral CONICET de la disciplina Ciencias de la Computación e Informática y un aspirante a becas EVC-CIN.

2. Motivación

En Julio de 2020 se relanzó la iniciativa Pampa Azul [4], destinada a promover el conocimiento científico, el desarrollo tecnológico y la innovación productiva en el Atlántico Sur, con el fin de crear una cultura del mar en la sociedad argentina, fomentar el uso sostenible de los bienes naturales marinos y fortalecer el crecimiento de la industria nacional asociada.

Sin embargo, el manejo en línea de los datos generados en el primer lanzamiento de Pampa Azul (año 2014) dejó en claro que la gestión y modelado de los datos en línea necesitaba ser planificado, resguardado y compartido para que los productos que se generen con ellos puedan ser utilizados por la comunidad científica para una adecuada comprensión del funcionamiento de nuestros espacios marinos. Más aún, no se tuvo en cuenta la integración de estos datos con bases federadas de alcance mundial, por lo cual se hace muy dificultoso el aprovechamiento científico integral de los mismos.

3. Líneas de Investigación y Desarrollo

Este proyecto desarrolla como principal línea de investigación, el Modelado Conceptual en la Web Semántica [5] y la construcción de grafos de conocimiento oceanográfico [6, 7] mediante datos enlazados [8, 9] para la integración de datos científicos y su explotación [10, 11]. Pero además, será necesario abordar el tratamiento de grandes volúmenes de datos oceanográficos y meteorológicos [12], caracterizados por las 5V (Volumen, Velocidad, Variedad, Verosimilitud y Valor) de Big-data [13, 14] y teorías, técnicas y herramientas para la

visualización de datos científicos oceanográficos [15, 16].

4. Resultados esperados

El objetivo general del proyecto es estudiar y desarrollar grafos de conocimiento para la gestión integrada y visualización de datos en línea de ciencias del mar a través de tecnologías Big Data integrando datos provenientes de campañas oceanográficas, de repositorios de Biodiversidad marina y de datos ambientales. En una primera etapa será acotado al golfo San Jorge (GSJ) dado que se cuenta con datos provenientes de las campañas realizadas por el grupo de trabajo GSJ, 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. El impacto esperado de esta investigación será el de permitir una gestión confiable de los datos en línea, para garantizar la preservación y el acceso a nuestros activos nacionales de investigación multidisciplinar.

Se esperan como resultados:

1. Extender el modelo conceptual BiGe-Onto [17] y formalizar el grafo de conocimiento OceanGraph [18] desarrollado durante el proyecto UNPSJB-PI 1562, incluyendo nuevos registros que agregan complejidad a la información.

2. Estudio y desarrollo de una plataforma de visualización en línea para proveer facilidades de analíticos visuales y permitir consultas y análisis interactivos de la información proveniente de los grafos de conocimiento.

3. Estudiar e implementar sistemas de razonamiento que permitan la explotación de conocimiento implícito en el grafo.

4. Implementar el plan de gestión de datos que describe la planificación, organización y previsión de la colección digital de los datos recolectados, reutilizados o procesados. Para su construcción se toma como guía el Plan de Gestión de Datos CONICET.

5. Formación de recursos humanos

En este proyecto participa un docente de UNPSJB-Puerto Madryn, doctor en Ciencias de la Computación y actual becario posdoctoral CONICET focalizado en el desarrollo de Grafos de Conocimiento para la gestión integrada de datos científicos multidisciplinares de ciencias oceanográficas. También participa un becario doctoral CONICET perteneciente a CESIMAR-CONICET que investiga analíticos visuales para datos enlazados en Ciencias del Mar. Sus directores son investigadores de la UNPSJB, de la Universidad Nacional del Sur y del CESIMAR-CENPAT-CONICET, quienes también integran el equipo de trabajo en calidad de asesores y expertos de dominio.

Además, el equipo de trabajo incluye a tres docentes del Departamento de Informática de la Facultad de Ingeniería de la UNPSJB-Puerto Madryn, que desarrollan líneas de investigación referidas a la Gestión de Información Científica y Tecnológica.

Referencias

[1] 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.

[2] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6, 06/2011 2011.

[3] Schlitzer, R. (2002). Interactive analysis and visualization of geoscience data with Ocean Data View. Computers & geosciences, 28(10), 1211-1218.

[4] Pampa Azul. https://www.argentina.gob.ar/noticias/se-relanzo-la-iniciativa-pampa-azul , 2020. [Online; accessed 22-Feb-2022].

[5] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American, 284(5):34–43, May 2001.

[6] 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. *In Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW 2018)*, 2018.

[7] Marcos Zárate, Pablo Rosales, Germán Braun, Mirtha Lewis, Pablo Rubén Fillottrani, and Claudio Delrieux. Oceangraph: Some initial steps toward a oceanographic knowledge graph. *In Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 33–40. Springer, 2019.

[8] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, II Vardeman, et al. Five stars of linked data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.

[9] 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.

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

[11] Giuseppe Andronico, Valeria Ardizzone, Roberto Barbera, Bruce Becker, Riccardo Bruno, Antonio Calanducci, Diego Carvalho, Leandro Ciuffo, Marco Fargetta, Emidio Giorgio, Giuseppe Rocca, Alberto Masoni, Marco Paganoni, Federico Ruggieri, and Diego Scardaci. e-infrastructures for e-science: A global view. *Journal of Grid Computing*, 9:155– 184, 06 2011.

[12] Adam Leadbetter, Robert Arko, Cynthia Chandler, Adam Shepherd, and Roy Lowry. Linked Data An Oceanographic Perspective. *The Journal of ocean Technology*, 8(3), 2013.

[13] Mark A Beyer and Douglas Laney. The importance of 'big data': a definition. *Stamford, CT: Gartner*, pages 2014–2018, 2012.

[14] Nikos Bikakis and Timos Sellis. Exploration and visualization in the web of big linked data: A survey of the state of the art. *arXiv preprint arXiv*:1601.08059, 2016.

[15] 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.

[16] Peter Fox and James Hendler. Changing the equation on scientific data visualization. *Science*, 331(6018):705–708, 2011.

[17] Zárate, M., Braun, G., Fillottrani, P., Delrieux, C., & Lewis, M. (2020). BiGe-Onto: an ontology-based system for managing biodiversity and biogeography data. Applied Ontology, 15(4), 411-437.

[18] Zárate, M., Rosales, P., Braun, G., Lewis, M., Fillottrani, P. R., & Delrieux, C. (2019). OceanGraph: Some initial steps toward a oceanographic knowledge graph. In Knowledge Graphs and Semantic Web (pp. 33–40). Springer International Publishing. https://doi.org/10.1007/978-3-030-21395-4_3