

Tecnologías de la Información facilitadoras para Linked Open Data (LOD)

Roxana Martínez, Pablo Vilaboa, Lucas Piersanti, Cecilia Taboada, Felipe Parrondo, Agustín Motta, Lucas Pascual, Gonzalo Pereyra

Instituto de Tecnología (INTEC)
Universidad Argentina de la Empresa (UADE)
Lima 775, Ciudad Autónoma de Buenos Aires, Argentina

{mariarmartinez, pvilaboa, lpiersanti, ctaboada, fparrondo, agmotta, lucapascual, gpereyrametnik}@uade.edu.ar

RESUMEN

Cada vez son más las personas que utilizan las Tecnologías de la Información como el medio de intercambio de datos, mediante plataformas de software con una infraestructura adecuada entre éstas. Esto permite que exista una mayor interoperabilidad de los datos que son compartidos y disponibilizados en un contexto de participación y colaboración de los ciudadanos y organismos tanto públicos, como privados. Linked Open Data (LOD) se refiere a datos publicados y enlazados mediante una determinada estructura web. Este tipo de enfoque permite conectar los datos independientemente de dónde estos residan, por lo que promueve la publicación y el enlace de datos en línea de una manera que los datos estén disponibles de forma abierta y accesible, y que puedan ser interconectados y consultados de manera efectiva. Algunos de los principios clave de Linked Open Data se relacionan con la disponibilidad de los datos, como así también su accesibilidad en la que deben estar disponibles en formatos abiertos y estándar, como RDF (Resource Description Framework), para que las máquinas puedan procesarlos.

Este trabajo utilizará una variedad de campos, desde la investigación científica hasta la gestión de datos gubernamentales y la creación de pequeños prototipos en línea. Además, permitirá conocer el estado situación actual de las tecnologías utilizadas en el contexto de datos abiertos públicos

vinculados, Linked Open Data (LOD). Finalmente, se presentarán las mejores prácticas en aspectos de niveles técnicos para el tratamiento de este tipo de datos enlazados en los sitios web de los portales gubernamentales.

Palabras clave: Linked Open Data (LOD), Tecnologías de la Información, RDF, Diseño y Arquitectura de componentes para datos vinculados.

CONTEXTO

El presente trabajo es parte del proyecto denominado “*Tecnologías de la Información facilitadoras para Linked Open Data (LOD)*”, que tuvo inicio en el mes de diciembre 2023. Este proyecto pertenece a la línea de investigación de Ingeniería de Software (IS) del Instituto de Tecnología (INTEC). Las actividades del Instituto están directamente relacionadas con las carreras grado y posgrado de la Facultad de Ingeniería y Ciencias Exactas de la Universidad Argentina de la Empresa (UADE). Alguna de las tareas de este proyecto, se realizan en los laboratorios disponibles en los UADE Labs, edificio tecnológico inaugurado en 2010. El proyecto es financiado y evaluado por la Coordinación de Investigaciones, unidad que depende de la Secretaría Académica de la Universidad, tiene una duración de 2 años, y cuenta con la participación de docentes y estudiantes de grado y posgrado en diversas carreras.

1. INTRODUCCIÓN

En la era digital y con el progreso tecnológico de Internet, la responsabilidad de los gobiernos de fomentar y promover iniciativas de datos abiertos ha aumentado en los últimos años. El intercambio de información se ha convertido en una herramienta fundamental para impulsar y promover los tres principios fundamentales del gobierno abierto: transparencia, participación y colaboración. [1]. El enfoque de Gobierno Abierto es un aspecto que fomenta la participación ciudadana y la colaboración en la toma de decisiones. Esta idea promueve la creación de una plataforma para la participación y colaboración ciudadana, con el objetivo de co-gestionar los servicios públicos [2]. Esto permite que los gobiernos sean más accesibles para la ciudadanía y que se tenga un mayor control sobre sus acciones. Como resultado de estas políticas, surgen los datos abiertos. Este término se refiere a los datos que se ponen a disposición en línea, los cuales deben ser reutilizables y redistribuibles por cualquier persona [3]. Los ejemplos más comunes de datos abiertos incluyen datos científicos, información gubernamental y conjuntos de datos relacionados con los servicios proporcionados por el estado.

La calidad en la elaboración de los archivos, conjuntos de datos, o mejor conocidos como dataset de datos abiertos, no solo garantiza su comprensión, sino que también, se refiere a una estrategia para publicarlos y vincularlos. Por otra parte, estos datos abiertos vinculados, se refieren a la apertura de éstos (lograr interoperabilidad) y, además, a su interconexión con otros conjuntos de datos. Esta última acción facilita la integración y reutilización en diversos entornos [4]. Por lo tanto, Linked Open Data (LOD) que es un conjunto de estándares web, facilita el acceso a los datos publicados abiertos.

En la actualidad, los principios de datos vinculados se aplican de manera frecuente, lo cual da lugar a lo que se conoce como la nube de datos abiertos vinculados [5]. En el enfoque del entorno de los datos abiertos

vinculados (LOD), se promueven herramientas para proporcionar más información sobre las relaciones que existen entre los datos guardados/persistentes. El tipo de relación que existe entre éstos puede ser útil frente a varios fines posibles, en especial, a la mejora del rendimiento del acceso a la obtención de una determinada información que puede ser aprovechada para la toma de decisiones [6]. Por otra parte, se puede afirmar que el valor de los datos aumenta con la utilización de las interconexiones del metadato, esto es, trabajar con LOD no solo vincula los datos en la web [7], sino que, además, al interconectarlos con diferentes tipos de fuentes, cobra un valor significativo [8] a la hora de analizar un escenario de caso de estudio.

Uno de los puntos de este trabajo de investigación, implica realizar foco en un framework que es conocido en este contexto de datos vinculados, este se llama RDF. Éste es el Framework de Descripción de Recursos (RDF) [9], el cual presenta un marco para la representación de información en la WEB. Este entorno, define una sintaxis (modelo de datos) que sirve para vincular todos los lenguajes y especificaciones basados en RDF. La estructura principal conlleva a un conjunto representado por el trinomio “sujeto-predicado-Objeto”, que se clasifica en tres tipos de nodos: IRI, literales y nodos en blanco.

En el contexto de la explicación de la estructura de RDF, es posible indicar que, RDF se centra en la tríada que consta de: a) Un *sujeto*; b) Un *predicado*; c) Un *objeto*.

El *sujeto* representa el recurso principal, el *predicado* indica la relación entre el sujeto y el objeto, y el *objeto* es el valor o recurso asociado. Esta estructura posibilita la modelación de un amplio conjunto de relaciones.

En la Figura 1, se muestra un ejemplo de RDF con sus distintos elementos.

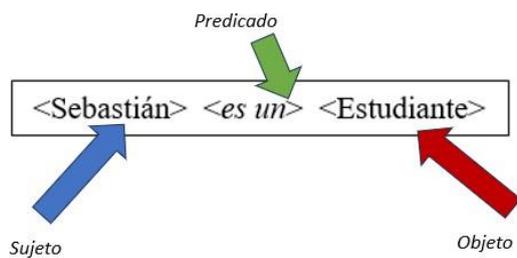


Figura 1. Ejemplo de RDF.

Básicamente, la estructura RDF, se puede entender como un mapa conceptual donde cada nodo se relaciona con un predicado.

La propuesta de este proyecto es aprovechar las distintas tecnologías de información para utilizarlas en el entorno de datos vinculados, y analizar diversas técnicas que nos permitan mejorar el análisis de casos de estudio para la toma de decisiones y valor agregado. Debido a que RDF es una propuesta basada en archivos XML, y archivos de validación como esquemas (XSD), es posible utilizar datasets de datos abiertos públicos que se encuentren en este tipo de formato con el fin de llevar a cabo, una representación de estructuras de tres partes “*sujeto-predicado-objeto*”, para mejorar el análisis de relación de datos y así lograr facilitar la correlación y el análisis de datos de múltiples conjuntos para descubrir patrones, tendencias o insights más profundos (entendimiento, percepción o conocimiento).

Otro punto para destacar de esta propuesta es que la combinación de los archivos XML (formatos del tipo abiertos para conjuntos de datos públicos abiertos), junto con las verificaciones de datos, ofrece la oportunidad de representar datos legibles por una computadora, lo que brinda una mayor interoperabilidad de software, como así también, mostrar la estructura de los datos de forma clara y estructurada. En este sentido, RDF no solo es una herramienta para la representación de los datos, sino también una forma sencilla de expresar el significado de cada uno de esos datos. Además, la información se interpreta de manera comprensible y legible. La fortaleza de RDF en la representación de relaciones semánticas y en la vinculación de los datos, contribuye a

la construcción de una red de datos interconectados en continuo crecimiento.

Finalmente, es importante concluir con que RDF proporciona un lenguaje simple para describir anotaciones sobre recursos web identificados por una URI (Uniform Resource Identifier), siendo esta una cadena de caracteres que contiene un nombre o una dirección que identifica a un objeto en la web [10]. Además, el foco de recurso se puede hacer referencia, a un sitio web, o bien un servicio web, entre otros.

Como conclusión, este proyecto considera algunos aspectos técnicos que brindan conceptos sobre diversas aplicaciones en el ámbito de la web semántica y la representación de datos estructurados [11]. Algunas de las aplicaciones más relevantes en las que se podrían hacer mayor hincapié de son las siguientes: a) Integración de Datos; b) Análisis de Aplicaciones; c) Interoperabilidad entre aplicaciones; d) Búsqueda y recuperación de información;

Por todo esto, es que este proyecto se centrará en el estudio de una selección de vocabulario en base a un dominio de determinados datos, como así también la limpieza y el armado de la muestra a trabajar. Posteriormente, se analizará la propuesta de reglas que darán origen a la generación de los datos RDF. Siendo una pieza fundamental, establecer una conexión entre los conceptos de datos abiertos, datos vinculados y RDF, ya que, gracias a éstos, se podrá complementar un mejor análisis en el ámbito de la gestión de datos en la web, y así lograr un valor agregado al conocimiento de lo trabajado.

2. LÍNEAS DE INVESTIGACIÓN DESARROLLO

Este proyecto pertenece a la línea de investigación de Ingeniería de Software (IS) del Instituto de Tecnología (INTEC). Los ejes principales del tema que se están investigando en referencia a las actividades de I+D+i (Investigación, Desarrollo e Innovación) son:

- Analizar las falencias en cuestiones tecnológicas actuales de vinculación de datos públicos.

- Elaborar el estado situación de las tecnologías involucradas.
- Desarrollar una guía de las mejores prácticas en aspectos de niveles técnicos para el tratamiento de los datos públicos vinculados que son asociados (relacionados entre sí) y visualizados en los portales estatales nacionales.
- Diseñar y desarrollar una propuesta de prototipo de software utilizando las mejores técnicas de diseño y arquitectura de software orientadas es este contexto.

Fundamentación del proyecto:

Desde el enfoque en Audiencia Científica:

Los resultados (parciales y finales) de este proyecto serán transferidos para contribuir con nuevo conocimiento a la comunidad académica y, a la sociedad en general en el ámbito nacional e internacional. Estas contribuciones se encontrarán plasmadas en forma de publicaciones.

Desde el enfoque en Audiencia como

Institución: La audiencia interna la constituyen los estudiantes y docentes de la carrera de Ingeniería Informática y afines, pudiendo también incorporarse integrantes de diferentes carreras interesados en las cuestiones del gobierno abierto en esta temática de datos vinculados.

Desde el enfoque Social: El proyecto corresponde a un abordaje desde la disciplina del Software. Su base fundamental surge por la necesidad de tener un mejor acercamiento y entendimiento de datos públicos, tanto con el ciudadano como con las organizaciones a través del uso de las TIC integradas y vinculadas.

3. RESULTADOS OBTENIDOS/ESPERADOS

Enfoque del proyecto:

Se buscará brindar nuevos enfoques especialmente en el diseño y construcción de prototipos de software para la validación y verificación en la vinculación de datos abiertos públicos a través de conceptos del tipo Linked Open Data (LOD). Por las características del proyecto guarda fuerte

vínculo con conceptos de Ingeniería de Software, Diseño de Software y Arquitectura de Software.

Objetivos principales:

Identificar los principales aspectos tecnológicos que son facilitadores para llevar a cabo el tratamiento de los datos abiertos públicos vinculados (LOD), y brindar buenas prácticas a niveles técnicos en lo que concierne a esta temática. Esto permitirá comprender los conceptos que son fundamentales en este nuevo paradigma gubernamental desde un aspecto de diseño y arquitectura de software.

Objetivos específicos esperados:

Relevar los conceptos fundamentales en esta temática que se relacionan con el marco de metadatos vinculados; Analizar y comparar los antecedentes a nivel internacional sobre dicho tema; Elaborar el estado situación del contexto de datos abiertos públicos vinculados; Comprender e identificar el marco regulatorio legal a nivel nacional; Relevar las distintas tecnologías a nivel técnico; Estudiar las técnicas más relevantes en contextos de enlace de datos en línea de una manera que los datos estén disponibles de forma abierta y accesible, y que puedan ser interconectados y consultados de manera efectiva; Definir las ventajas y desventajas de la implementación; Identificar las mejores prácticas en aspectos de niveles técnicos para el tratamiento de los datos públicos que son asociados (relacionados entre sí) y visualizados en los portales estatales nacionales. Esto permitirá crear una web de datos enlazados en la que las personas y algoritmos puedan acceder y utilizar información de diversas fuentes de manera más eficiente y significativa.

Metodología y Técnicas:

En los que respecta a las técnicas de investigación, se realizarán enfoques cualitativos que permitan obtener estudios enfocados a:

- Observación de contextos en esta temática.
- Investigación bibliográfica: Con el fin de

apoyar y sustentar el trabajo investigativo. Relevar los estudios previos y contrastarlos a la actualidad.

- Estudios de Casos en la República Argentina y países extranjeros: Para recolectar información pertinente al estudio y obtener una representación lo más similar a la realidad relevada.
- Estudio de Técnicas de LOD (como, por ejemplo, RDF: modelo conceptual para proporcionar información descriptiva sobre los recursos que se encuentran en la Web, permitiendo el intercambio de información a través de diferentes).
- Análisis y encuadre del framework de la metadata en datos abiertos públicos gubernamentales vinculados.

4. FORMACIÓN DE RECURSOS HUMANOS

Este proyecto se compone por 2 (dos) docentes de grado que cuentan con varios años de experiencia en el dictado de clases en varias Universidades de Argentina y más de 20 años de experiencia en el ambiente laboral en Tecnología Informática, también tienen estudios de posgrado: Uno de ellos es docente investigadora, Doctora en Ciencias Informáticas de la Universidad Nacional de La Plata (UNLP), actual directora del Proyecto de Investigación y Magister en Tecnología Informática de la Universidad Abierta Interamericana (UAI), además, el proyecto cuenta con otro docente investigador, Maestrando en Tecnología Informática de la Universidad Abierta Interamericana (UAI).

El equipo también cuenta con la participación de estudiantes de grado de las carreras de la rama de Informática y Sistemas de la UADE (Universidad Argentina de la Empresa), siendo éstas: Ingeniería Informática y Tecnicatura de Desarrollo de Software, su participación es en calidad de colaboradores en el proyecto. En relación directa con la línea de I+D+i presentada para el proyecto, 2 miembros del equipo se encuentran en realización de: 2 trabajos finales de carreras afines en la UADE.

5. BIBLIOGRAFÍA

- [1] Martínez, M. R. (2022). Métricas de calidad para validar los conjuntos de datos abiertos públicos gubernamentales (Doctoral dissertation, Universidad Nacional de La Plata).
- [2] Villodre, J. (2019). Innovación pública abierta. EUNOMÍA. Revista en Cultura de la Legalidad, (17), 314-327.
- [3] i Martínez, A. C. (2014). Los principios de los datos abiertos en la legislación española. IDP: revista de Internet, derecho y política= revista d'Internet, dret i política, (19), 62-77.
- [4] datos.gob.es (2017). "Linked data como modelo de datos". Disponible en: <https://datos.gob.es/es/noticia/linked-data-como-modelo-de-datos>
- [5] Hausenblas, M. (2011, May). Utilising linked open data in applications. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics (pp. 1-4).
- [6] Yang, H. C., & Hsu, C. C. (2015). Semantic Recommendation Using Linked Open Data. In Proceedings of the ASE BigData & SocialInformatics 2015 (pp. 1-4).
- [7] Azad, H. K., Deepak, A., & Abhishek, K. (2016, March). Linked open data search engine. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (pp. 1-5).
- [8] Ávila-Barrientos, E. (2022). Recuperación de información con Linked Open Data. Investigación bibliotecológica, 36(91), 125-146.
- [9] Klune, G., Carroll, J., & McBride, B. (2014). w3c. Disponible en: <https://www.w3.org/TR/rdf11-concepts/>
- [10] Acevedo, S., & Muñoz, O. (2022). Framework para determinar el contexto de los datos basado en ontologías de dominio.
- [11] IBM (2021). Creación de documentos de vocabulario de RDF y asignación de URI.