Conferencia Internacional BIREDIAL – ISTEC 2018

VIII Conferencia Internacional sobre Bibliotecas y Repositorios Digitales de América Latina

Eje temático: Ciencia Abierta – Datos Abiertos de Investigación

Tipo de Trabajo: Ponencia

Título: ***Integración de Repositorios Semánticos, un camino hacia los Datos Abiertos Enlazados***

Autores: *Carlos Buckle[1,2], Marcos Zarate[1,2,4], Gustavo Samec[1,2,3], Renato Mazzanti[1,2,3]*

Afiliaciones:

(1) Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Boulevard Brown 3051, Puerto Madryn, Chubut, Argentina.

(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-CENPAT-CONICET)

Contacto para la ponencia: *Renato Mazzanti - Celular +54-9-2804193476*

Requerimientos para la presentación de la ponencia: *Computadora y Proyector*

Dirección Postal: *Boulevard Brown 3051, (9120) Puerto Madryn, Chubut, Argentina.*

Teléfonos Institucionales UNPSJB Puerto Madryn: *(+54 280) 4883585 / 4883499*

Datos y Referencias Biográficas de los Autores: *(Anexas al final del trabajo)*

Integración de Repositorios Semánticos un camino hacia los Datos Abiertos Enlazados

Carlos Buckle, Marcos Zárate, Gustavo Samec, Renato Mazzanti

Resumen:

La Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Argentina, se encuentra en proceso de implementación de su Repositorio Digital Institucional de Acceso Abierto (RDI). En este contexto, se ha formulado un proyecto de investigación "*Infraestructura de Acceso a Datos Primarios con aporte de semántica en Repositorios Digitales*" que entre sus objetivos propone desarrollar modelos y componentes para integrar las producciones científicas residentes en el RDI con los datos primarios que las sustentan. Dicho proyecto propone el uso de tecnologías de la web semántica y estándares de la World Wide Web Consortium (W3C) para la definición de consultas integradas entre repositorios que faciliten el descubrimiento de conocimiento a los investigadores.

En este trabajo se exponen los avances de la etapa inicial del proyecto, que consiste en enlazar publicaciones científicas con datos primarios citados en ellas utilizando Resource Description Framework (RDF). Se desarrolló una experiencia que vincula los metadatos de publicaciones científicas de DSpace con datasets del Sistema Nacional de Datos Biológicos (SNDB). La propuesta se valida mediante consultas integradas en Protocol and RDF Query Language (SPARQL), demostrando a su vez, las potencialidades de explotación semántica. Estos resultados permiten guiar el proyecto hacia la publicación de datos abiertos enlazados (Linked Open Data, LOD), que se presentan como el camino estándar hacia la integración de datos abiertos globales entre diferentes dominios.

Palabras Clave: Repositorio Semántico, Datos Abiertos Enlazados, Interoperabilidad Semántica, Datos científicos primarios

Abstract:

National University of Patagonia San Juan Bosco (UNPSJB), Argentina, is in the process of develop its Digital Institutional Repository (DIR). In this context, a research project called "Access Infrastructure to Primary Data with Semantic Contribution in Digital Repositories" has been formulated, among its objectives proposes to develop models and components to integrate the scientific productions stored in DIR with the underlying primary data. This project proposes the use of semantic web technologies and World Wide Web Consortium (W3C) standards to define integrated queries between repositories to simplify the discovery of knowledge to researchers.

In this paper we present the results of the initial stage of the project, which consists of linking scientific publications with primary data cited therein, to accomplish this, the information was stored in Resource Description Framework (RDF). An

experience was developed linking the metadata of scientific publications of DSpace with datasets from the National Biological Data System (SNDB) previously converted to RDF. The proposal is validated through integrated queries using Protocol and RDF Query Language (SPARQL), demonstrating the potential of semantic exploitation. These results allow conduct the project towards the publication of Linked Open Data (LOD), which is presented as the standard way towards the integration of global open data between different domains.

Keywords: Semantic Repositories, Linked Open Data, Semantic Interoperability, Primary Data in Science

1. Introducción

La Ley 26.899[1] de Repositorios Digitales Institucionales de Acceso Abierto (RDI), establece la obligatoriedad de desarrollar repositorios digitales, propios o compartidos, por parte de los organismos e instituciones públicas que componen el Sistema Nacional de Ciencia, Tecnología e Innovación y que reciben financiamiento del Estado Nacional. Además de proveer acceso abierto a la producción científico-tecnológica documental, generada en artículos de revistas, libros, trabajos de congresos, reportes, tesis, etc. la ley requiere el establecimiento de políticas institucionales para la gestión, el acceso público y la preservación de datos primarios de investigación, garantizando que sean de acceso libre y compatibles con las normas de interoperabilidad adoptadas internacionalmente.
En la reglamentación de la ley[2], se establece el Sistema Nacional de Repositorios Digitales (SNRD) como instrumento técnico-operativo para nuclear e impulsar a los RDI nacionales y se lanzan emprendimientos para la conservación y publicación de datos primarios de investigación como el Programa de Grandes Instrumentos y Bases de Datos, el cual alberga actualmente al Sistema Nacional de Datos Biológicos (SNDB) y al Sistema Nacional de Datos del Mar (SNDM), entre otros.

Adhiriendo a estas iniciativas, la Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB) reguló la creación de su propio RDI y con ese impulso se han generado un conjunto de acciones institucionales, entre ellas, la formulación de actividades de investigación aplicada. Uno de los casos es el proyecto "*Infraestructura de Acceso a Datos Primarios con aporte de semántica en Repositorios Digitales*" el cual pretende desarrollar herramientas que faciliten a los investigadores su tarea de búsqueda, análisis, validación y descubrimiento de conocimiento existente en repositorios públicos de datos primarios. Para ello propone un desarrollo experimental que aportará modelos y componentes enfocados principalmente a facilitar el acceso a datos primarios de manera natural y con lenguaje unificado, como así también a mantener el vínculo entre las producciones documentales y los conjuntos de datos que las sustentan a través de

[1] Ley 26.899. "Repositorios digitales institucionales de acceso abierto". Sancionada: Nov 2013. Promulgada: Dic 2013. Boletín Oficial de la República Argentina. Presidencia de la Nación.
[2] Resolución 753/16, Minist. de Ciencia, Tecnolog a e Innovación Productiva. Argentina. Nov 2016

citas a datos (data citation). Una motivación para esto se origina en la realidad global donde la mayoría de los RDIs implementados alberga grandes volúmenes de material documental pero un porcentaje muy bajo de datos científicos vinculados (Arano et al., 2011).

El desafío significa resolver aspectos de interoperabilidad a nivel semántico y uniformidad de vocabularios, para lo cual es necesario repensar los modelos clásicos de organización del conocimiento en RDIs. Es allí donde comienzan a adquirir relevancia las teorías y tecnologías de la web semántica (Berners-Lee, Hendler, & Lassila, 2001), como así también datos abiertos enlazados (LOD - Linked Open Data) (Janowicz et al., 2014), los cuales se presentan como el camino estándar hacia la integración de datos entre repositorios abiertos heterogéneos y de diferentes dominios (Smith et al., 2003).

Este trabajo, como un tránsito hacia los LOD, presenta una experiencia de integración entre un RDI y un repositorio de datos primarios vinculadas a través de citas de datos, sobre las cuales se realizan consultas semánticas integradas utilizando SPARQL[3], previa generación de repositorios RDF[4] y respetando la metodología utilizada para la conversión y publicación establecida en (Hyland, Atemezing, & Villazón-Terrazas, 2014).

La infraestructura de RDI que se utiliza es DSpace (Smith et al., 2003), que es la seleccionada para el RDI de la UNPSJB y la base de datos primarios a la cual se hace referencia corresponde a un conjunto de colecciones biológicas gestionada en CONICET-CENPAT[5] y que forma parte del SNDB.

2. Fundamentos para la Experiencia

En esta sección se presentan aspectos conceptuales que resultan fundamentales como basamento para la experiencia. En primer lugar: LOD (Heath & Bizer, 2011) como propuesta para la gestión de datos abiertos publicados y vinculados en repositorios desde la perspectiva de la Web Semántica, y en segundo lugar: Citas a Datos Abiertos, para cubrir la necesidad de vinculación entre RDI y fuentes de datos primarios de investigación.

2.1. LOD - Linked Open Data

Permite publicar y conectar datos estructurados en la web, de manera que sean comprensibles por las máquinas y sean autodescriptos con significado explícitamente definido. Estos datos, vinculados con otros, generan en la web, una colección de tripletas RDF a las que se hace referencia mediante URI (uniform resource identifiers) en los diferentes espacios de nombres. Estas capacidades son fundamentales para la implementación de la Web Semántica.

[3] https://www.w3.org/TR/rdf-sparql-query/
[4] https://www.w3.org/RDF/
[5] Centro Científico Tecnológico CENPAT Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) - Puerto Madryn - Argentina http://www.cenpat-conicet.gob.ar/

El uso de LOD permite:

- Integración y Reusabilidad: Elimina la barrera de los silos de datos. Pues no necesita de protocolos propietarios o interfaces de aplicaciones (Aplication Program Interface, APIs), de terceros para recuperar información de diferentes repositorios. Permite de esta forma integrar información heterogénea sin depender de terceros.
- Semántica: los datos dejan de ser ambiguos y pueden ser interpretados y entendidos tanto por seres humanos como por otras aplicaciones de software.
- Visibilidad y fácil detección del conjunto de datos.
- Expresividad: dada por los grafos RDF que describen información.
- Consultas: mediante SPARQL, que permite un mayor alcance, búsquedas por significado y posibilidad de recuperar datos desde diferentes repositorios.
- Facilidad de búsquedas: un usuario puede realizar búsquedas en varios repositorios, desde un end-point SPARQL. También puede descargar parte de los datos y combinarlos con otros datos y procesos de acuerdo a sus necesidades.

Las buenas prácticas en el tratamiento de LOD, impulsan a la adopción de una metodología e indican que el proceso de publicación debe respetar un ciclo de vida, iterativo e incremental, que garantice la mejora continua tanto desde puntos de vistas cualitativos como cuantitativos. El ciclo de vida propuesto en (Hyland et al., 2014) se muestra en Fig. 1 y propone las siguientes etapas:

- *Especificar*: incluye el diseño de URIs, definir y describir las fuentes de datos.
- *Modelar*: Buscar ontologías adecuadas, con vocabularios que sean apropiados de acuerdo a las fuentes de datos y crear el modelo en base a la reutilización de los modelos seleccionados.
- *Generar*: Transformar la fuente de datos a RDF, depurar los datos y vincularlos con otros conjuntos de datos.
- *Publicar*: publicar los datos generados y habilitar consultas.
- *Explotar*: Hacer uso de los datos y aplicaciones que en base a consultas faciliten el descubrimiento y extracción de información.

Fig. 1. Ciclo de Vida de LOD

En 2006 Berners-Lee definió cuatro reglas para la publicación de LOD[6]:

- Usar URIs identificando los recursos de forma unívoca.
- Usar URIs http para que sea posible acceder a la información del recurso.
- Ofrecer información sobre los recursos usando RDF.
- Incluir enlaces a otros URIs.

[6] http://www.w3.org/DesignIssues/LinkedData.html

Y un ranking, basado en 5 estrellas, que permite evaluar la calidad de los LOD. Esto fue posteriormente ampliado en (Janowicz et al., 2014), que permite evaluar la calidad de los vocabularios utilizados en los LOD.

2.2. Citas a Datos Abiertos (Open Data Citation)

El impacto de la investigación científica continúa siendo evaluado por mecanismos que dependen de las tasas de citas de la literatura publicada, por lo que los investigadores, financiadores y organizaciones que realizan investigaciones están justificadamente interesados en aumentar las citas.
En el contexto de materiales impresos, como libros y revistas, la cita es bien conocida. No es así en el mundo digital, donde actualmente se desarrolla la producción científica, en el que se está recopilando una cantidad creciente de información en bases de datos curadas que evolucionan constantemente. Los propietarios de las bases de datos, los editores y los grupos de estándares deben considerar cómo se deben citar dichos datos. Una pregunta que surge es: "¿cómo podemos generar citas para consultas generales sobre la base de datos, es decir, aquellas que no corresponden a las visitas a la página web de la base de datos?"(Davidson, Buneman, Deutch, Milo, & Silvello, 2017).
A diferencia de las publicaciones tradicionales donde la granularidad es fija -ejemplo: acta de conferencia, paper de congreso, etc.- la granularidad del material a ser citado por estas consultas es variable. Potencialmente hay infinitas posibilidades de consultas que acceden a diferentes subconjuntos de registros de las bases de datos y no se puede esperar que se expliciten como citas la totalidad de estas consultas. En su lugar, debemos encontrar maneras de usar citas producto de consultas más generales, construidas automáticamente, para algunos subconjuntos de la base de datos. Esto se constituye en un problema computacional como argumenta (Buneman, Davidson, & Frew, 2016).

(Ball & Duke, 2011) Indican que los datos deben considerarse productos de investigación legítimos y válidos. Las citas de datos deben tener la misma importancia en el registro académico que las citas de otros objetos de investigación, como las publicaciones. Que las citas de datos deberían facilitar el otorgamiento de créditos académicos y la atribución normativa y legal a todos los contribuyentes de los datos, reconociendo que puede no ser aplicable a todos los datos un estilo o mecanismo único de atribución. Que en la literatura académica, siempre que demande datos, se deben citar los datos correspondientes. Que una cita de datos debe incluir un método persistente de identificación que sea accionable por la máquina, globalmente único y ampliamente utilizado por la comunidad. Que las citas de datos deberían facilitar el acceso a los mismos datos y a los metadatos, documentación, código y otros materiales asociados, según sean necesarios tanto para humanos como máquinas para usar la información de los datos referenciados. Que los identificadores únicos y los metadatos que describen los datos, y su disposición,

deben persistir, incluso más allá de la duración de la vida de los datos que describen. Que las citas de datos deberían facilitar la identificación, el acceso y la verificación de los datos específicos que la respaldan. Que los métodos de citación de datos deberían ser lo suficientemente flexibles para adaptarse a las prácticas variadas entre las comunidades, pero no deberían diferir tanto como para poner en peligro la interoperabilidad de las prácticas de citación de datos entre las comunidades.

La forma más habitual de referenciar los datos primarios en una publicación es a través de una declaración de acceso a datos.
Para los datos abiertos, esta declaración debe indicar: desde cuándo está disponible, en que repositorio y proporcionar una URL y un identificador o un código de acceso para ayudar a acceder a los datos.
Para los datos restringidos, la declaración debe indicar el motivo legal o ético de la restricción y proporcionar un enlace a un registro permanente que explique las condiciones de acceso.
Si bien una declaración simple de este tipo cumple con la necesidad básica de referencia de datos, no se cumple con varios aspectos:

- Si hay un error tipográfico en el identificador o URL, no hay información adicional para ubicar los datos entre las existencias del repositorio;
- Los autores pueden tener la tentación de dar la URL del repositorio, en lugar de una específica para el conjunto de datos;
- No da el debido crédito a los creadores del conjunto de datos, un punto especialmente importante si estos son diferentes de los autores de la publicación;
- No trata los datos como un registro de primera clase de investigación.

Todos estos problemas pueden resolverse mejorando la declaración con una cita de datos. Al igual que con otras citas, esto implica proporcionar un puntero en el texto a una entrada en la lista de referencias.
Si el editor no está dispuesto a aceptar una cita de datos, a veces es posible evitar esto citando un documento de datos, por ejemplo ZooKeys[7].
Los elementos que conformarían una cita de datos completa son un tema de debate (Altman & King, 2008) (Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011) (Green, 2009) (Starr & Gastl, 2011).

3. Infraestructura, Herramientas y Métodos utilizados para la experiencia

Como se comentó inicialmente, la experiencia presentada en este trabajo consistió en generar dos conjuntos de datos en RDF, uno con las publicaciones científicas del RDI de la UNPSJB (instancia de test) y otro con los datos primarios del SNDB citados en ellas, para luego validarlas y mostrar la potencialidad de explotación utilizando consultas semánticas integradas en SPARQL.
En este apartado se describen las herramientas utilizadas y las metodologías aplicadas. Un resumen gráfico de la infraestructura se puede ver en Fig. 2.

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4395841/

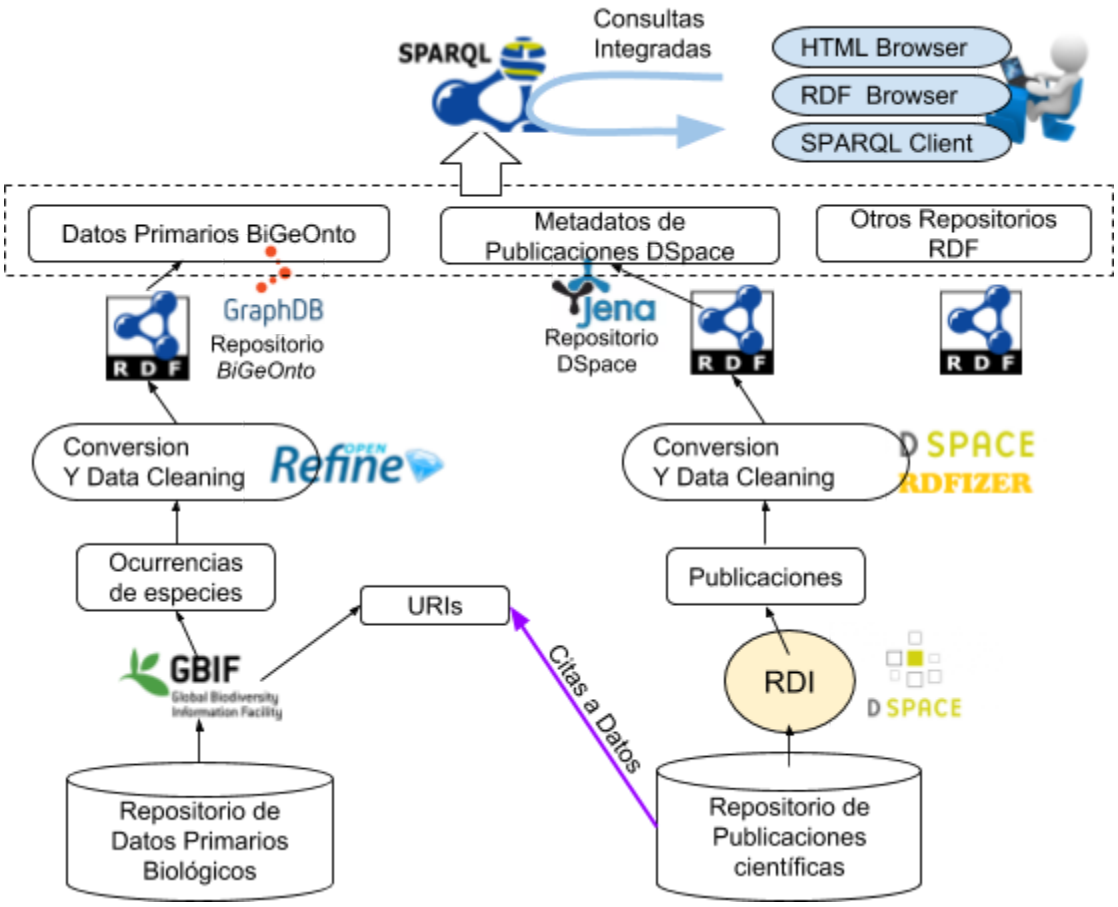

Fig. 2. Infraestructura utilizada para la experiencia

Para la experiencia se han utilizado las siguientes herramientas:

3.1. RDI

La UNPSJB seleccionó DSpace para la implementación de su RDI. Es un software de código abierto que provee herramientas para la administración de colecciones digitales. Actualmente es una de las aplicaciones más utilizada por universidades y organizaciones como repositorio institucional. El mismo da soporte a una gran variedad de datos, incluyendo libros, tesis, fotografías, videos, datos de investigación y otras formas de contenido.

DSpace organiza el repositorio en comunidades, colecciones e ítems. La comunidad es el nivel más alto de la jerarquía en el mismo, puede ser una facultad, un instituto, un centro de investigación, un laboratorio, etc. Una comunidad también puede tener sub-comunidades y dentro de cada comunidad contiene una o más colecciones. A su vez una colección puede contener uno o más ítems. Un ítem está compuesto por metadatos que describe su contenido y uno o más archivos que forman su contenido.

Para la descripción del contenido de un ítem se utilizan metadatos del estándar Dublin Core[8], de los cuales algunos son obligatorios y los utiliza el mismo Dspace para su gestión. Por otro lado, permite incluir nuevos esquemas de metadatos (siempre se recomienda utilizar esquemas de metadatos estándares como BIBO[9], FOAF[10], etc.) para la descripción de los ítems almacenados y facilitar su búsqueda y recuperación.

Dado que el RDI de la UNPSJB aún no está en producción las pruebas se realizaron en instancia de Test, basada en una copia del RDI del CCT CONICET-CENPAT[11] que tiene una infraestructura equivalente.

3.2. Fuente de Datos Primarios

Como mencionamos anteriormente la fuente de datos primarios fue extraída del SNDB[12], el cual almacena una base de datos unificada de información biológica, a partir de datos taxonómicos, ecológicos, cartográficos, bibliográficos, etnobiológicos y otros temas afines. La infraestructura para el SNDB está soportada por Global Biodiversity Information Facility (GBIF)[13], una organización internacional para promover que los datos científicos sobre biodiversidad estén disponibles en Internet y publica sus conjuntos de datos a través de una plataforma conocida como Integrated Publishing Toolkit (IPT) (Robertson et al., 2014), que es una herramienta gratuita de código abierto utilizada para publicar y compartir conjuntos de datos de Biodiversidad. La plataforma IPT gestiona los contenidos en formatos no estructurados o semiestructurados, reduciendo la posibilidades de interoperar con otros conjuntos de datos o hacerlos accesibles para máquinas. Los datos están disponibles como Darwin Core Archives (DwC-A) (Remsen, 2011), que consiste en un conjunto de archivos para describir la estructura y relaciones de los datos primarios junto con los archivos de metadatos que utiliza el estándar Darwin Core (DwC) (Wieczorek et al., 2012). Sin embargo, aunque SNDB utiliza DwC como un vocabulario común, opera de forma aislada.

3.3. Repositorios RDF

Se han utilizado dos repositorios RDF. Uno para los datos primarios y otro para las publicaciones del RDI-UNPSJB.

- Para el almacenamiento de los conjuntos de datos de biodiversidad (datos primarios) se utilizó GraphDB[14] que admite diferentes serializaciones de RDF. GraphDB permite a los usuarios explorar visualmente la jerarquía de clases, donde se puede explorar cada clase para explorar sus instancias.

[8] http://dublincore.org/documents/dcmi-terms/
[9] http://bibliontology.com/
[10] http://xmlns.com/foaf/spec/
[11] http://tango.cenpat-conicet.gob.ar:8080/xmlui/
[12] http://datos.sndb.mincyt.gob.ar/
[13] https://www.gbif.org
[14] https://ontotext.com/products/graphdb/

Es importante tener en cuenta que también podemos importar datos a GraphDB desde otros SPARQL endpoints que permiten consultas federadas, por ejemplo, podemos utilizar conjuntos de datos reconocidos internacionalmente como el *Universal Protein Resource* (UniProt)[15] para complementar la información taxonómica de una determinada especie.

- Para almacenar el conjunto de metadatos de publicaciones del RDI se utilizó Jena Fuseki[16] que además de exponer las tripletas como un end-point SPARQL accesibles desde http, está integrado con el triple store TDB[17]. Todas estas herramientas pertenecen al proyecto Apache Jena[18] que desarrolla aplicaciones y APIs open source en Java para la Web Semántica y Linked Data.

En la experiencia se han aplicado las siguientes metodologías para la definición de citas de datos y para la extracción de datos a RDF:

3.4. Método de Cita de Datos en Dspace

Como se indicó anteriormente, un ítem en DSpace está compuesto por metadatos que describen su contenido y uno o más archivos que forman su contenido. Para esta experiencia se utilizó el medatado *dc.relation.uri*[19] del estándar Dublin Core en todos los ítems que hacen referencias a dataset de datos primarios.

3.5. Método para creación de RDF en base a Datasets

Para convertir los conjuntos de datos primarios de biodiversidad se utilizó la metodología propuesta en (Zárate, Braun, & Fillottrani, 2017) (Ver Fig. 3) la misma consiste en convertir los archivos DwC-A a RDF en las siguiente etapas:

- Extracción, limpieza y reconciliación de datos: Los archivos DwC-A se extraen manualmente de la plataforma IPT y los datos primarios (ocurrencias) se procesan previamente (limpieza, conversión de tipos de datos, eliminación de valores nulos, etc.) utilizando la herramienta OpenRefine[20]. Esta herramienta también permite agregar servicios de reconciliación basados en SPARQL endpoint, que devuelven candidatos que pertenecen a conjuntos de datos externos para la reconciliación con campos de nuestro conjunto de datos. Se utilizó el punto final de DBpedia para generar enlaces a especies taxonómicas equivalentes.
- Definición de las URIs: Después de la limpieza y reconciliación, los datos se convierten a tripletas RDF utilizando RDF Refine[21]. Aquí es donde definimos el esquemas RDF especificando el sujeto, el predicado y el objeto de las

[15] https://sparql.uniprot.org/sparql
[16] https://jena.apache.org/documentation/fuseki2
[17] https://jena.apache.org/documentation/tdb
[18] https://jena.apache.org
[19] http://dublincore.org/documents/dcmi-terms
[20] http://openrefine.org/
[21] https://github.com/fadmaa/grefine-rdf-extension/releases

tripletas que se generarán. El siguiente paso es configurar los prefijos para vocabularios conocidos como por ejemplo la ontología básica *W3C Geo*, *DBpedia, FOAF, y DwC*. Cada recurso debe tener una URI definida para otros recursos dentro de este conjunto de datos y otros en cualquier parte de la web puedan referenciarlo. La URI base común para todos los recursos que definimos es *http://www.cenpat-conicet.gob.ar/resource/*

- Publicación: Los datos transformados de biodiversidad se han publicado y se puede acceder a ellos a través de GraphDB.
- Explotación: Mediante consultas SPARQL implementadas en un endpoint de GraphDB (Ver Sección 4).

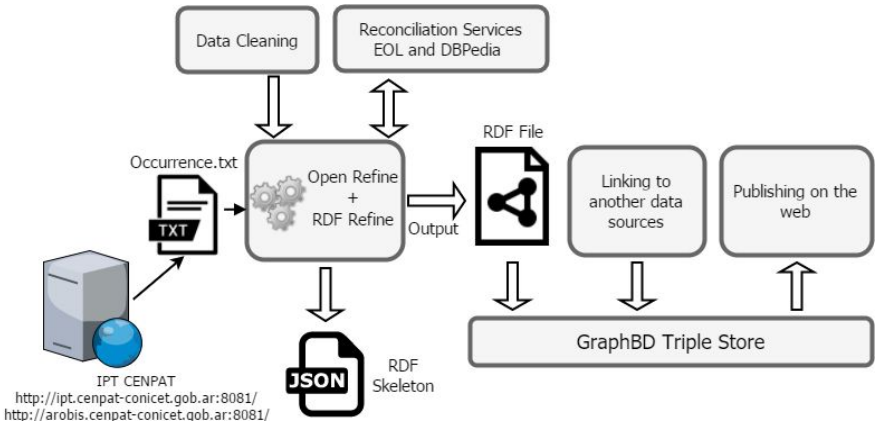

Fig. 3. Proceso de transformación utilizado convertir conjuntos de datos de biodiversidad

3.6. Método para creación de RDF en base a publicaciones en DSpace

Para convertir a RDF los metadatos de publicaciones almacenadas en el repositorio DSpace es necesario instalar un triple store, configurar y utilizar herramientas para realizar su exportación y por último habilitar a DSpace para que todo cambio realizado en sus datos se actualice en el triple store.

- Instalación de un Triple Store: DSpace puede utilizar cualquier triple store que soporte SPARQL 1.1 Query Language[22] y SPARQL 1.1 Graph Store HTTP Protocol[23]. Si bien SPARQL 1.1 permite modificar registros de un triple store, los mismos no se reflejaran en el repositorio en DSpace por lo que esta opción afectaría la integridad de los datos del mismo. A los efectos de realizar pruebas integradas entre dos triple store diferentes, para el caso de los metadatos de publicaciones almacenadas en DSpace, se utilizó Jena Fuseki 2.0.
- Exportar de los datos del Repositorio a RDF: Dentro de las aplicaciones que provee DSpace se instala una herramienta, conocida como RDFizer[24] en el ámbito de la Web Semántica, que exporta el contenido del repositorio en tripletas RDF. La misma permite exportar todo o parte del repositorio, borrar

[22] https://www.w3.org/TR/sparql11-query/
[23] https://www.w3.org/TR/sparql11-http-rdf-update/
[24] http://wiki.opensemanticframework.org/index.php/RDFizer_Concept

o actualizar una parte, etc. Esta herramienta sólo exporta los metadatos que describen las comunidades, colecciones e ítems y referencias a los archivos que contienen los ítems. DSpace provee una serie de archivos de configuración (escrito en Turtle) que le indica a esta herramienta como realizar el proceso.

4. Experiencia Realizada y Resultados obtenidos

Utilizando las técnicas, herramientas y metodologías comentadas en la sección anterior se realizó la experiencia de extracción de datos primarios y publicaciones a los correspondientes triple store RDF y se definieron consultas integradas con SPARQL. A continuación se muestran los resultados obtenidos:

A. *Triple Store de conjunto de datos primarios*: El usuario puede explorar visualmente el conjunto de datos, a través de la interfaz visual de GraphDB[25] para ello deberá utilizar las siguiente credenciales (usr: *bigeonto* pass: *bigeonto*) y luego seleccionar el repositorio denominado *BiGeOnto*. La Tabla 1 resume los principales enlaces para acceder y las estadísticas más relevantes.

Repositorio	BiGeOnto (user: *bigeonto* password: *bigeonto*)
URL	http://web.cenpat-conicet.gob.ar:7200/login
SPARQL Endpoint	http://web.cenpat-conicet.gob.ar:7200/repositories/BiGeOnto
SPARQL Endpoint Visual	http://web.cenpat-conicet.gob.ar:7200/sparql
Jerarquia de clases	http://web.cenpat-conicet.gob.ar:7200/hierarchy
Nro. de vocabularios	18
Nro. de clases	9
Nro. de propiedades	50
Nro. de Tripletas	4.334.668

Tabla 1: Principales características del Triple Store RDF de datos primarios

B. *Triple Store de conjunto de publicaciones*: El usuario puede explorar visualmente el conjunto de datos, a través de la interfaz visual de Jena Fuseki para ello seleccionar el repositorio denominado *DSpace*. La Tabla 2 resume los principales enlaces para acceder y las estadísticas más relevantes.

Repositorio	dspace
URL	http://tango.cenpat-conicet.gob.ar:3030
SPARQL Endpoint	http://tango.cenpat-conicet.gob.ar:3030/dspace/sparql
SPARQL Endpoint Visual	http://tango.cenpat-conicet.gob.ar:3030/dataset.html
Nro. de vocabularios	7

[25] http://web.cenpat-conicet.gob.ar:7200/login

Nro. de clases	1
Nro. de propiedades	28
Nro. de Tripletas	10.410

Tabla 2: Principales características del Triple Store RDF de publicaciones

C. *Definición de una Consulta Integrada SPARQL*: A los efectos de validar la experiencia y de mostrar la potencialidad de explotación de información se ha formulado una consulta SPARQL que requiere del acceso integrado a ambos triple store (datos primarios y publicaciones). La consulta formulada fue:

> "*Recuperar las publicaciones del RDI que han utilizado datos primarios recogidos en la ciudad de Puerto Madryn"*

La consulta expresada en SPARQL se puede ver a continuación y ejecutar en el siguiente link[26]

```
#Recuperar las publicaciones del RDI que han utilizado
#datosprimarios relacionados a Puerto Madryn, Chubut".

PREFIX dc: <http://purl.org/dc/terms/>
PREFIX bigeonto: <http://www.w3id.org/cenpat-gilia/bigeonto/>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX geo-pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX dcelem:<http://purl.org/dc/elements/1.1/>

SELECT ?titledspace ?publisher ?date
WHERE {
#Esta parte de la consulta recupera los DOIs de los conjuntos de
#datos primarios que tienen ocurrencias en Puerto Madryn.
    ?s a dwc:Occurrence.
    ?s bigeonto:memberOf ?dataset.
    ?dataset dc:identifier ?identi.
    ?s bigeonto:has_event ?event.
    ?event bigeonto:has_location ?location.
    ?location dwc:locality ?locality.
    # Filtramos por localidad
    FILTER(regex(str(?locality), "Puerto Madryn" ) ).
    # Filtramos solo los DOIs
    FILTER ( strstarts(str(?identi), "https://doi.org") ).

   # Invocamos al endopoint jena (respositorio DSpace)
   SERVICE <http://200.41.229.219:3030/dspace/sparql> {
       ?item dc:title ?titledspace.
       ?item dcelem:publisher ?publisher.
       ?item dc:issued ?date.
       ?item dc:relation ?url.
       # Filtramos los DOIs de las publicaciones
```

[26] http://web.cenpat-conicet.gob.ar:7200/sparql?savedQueryName=datos-primarios

```
        FILTER ( strstarts(str(?url), "https://doi.org") ).
        # Comparamos si existen DOIs iguales en ambos repositorios
        FILTER (?url = ?identi )
    }
}
GROUP BY ?titledspace ?publisher ?date
LIMIT 10
```

5. Conclusiones

En este trabajo presentamos la experiencia realizada y los resultados preliminares obtenidos al intentar resolver aspectos de interoperabilidad a nivel semántico y uniformidad de vocabularios entre repositorios documentales y datos primarios citados. Utilizando tecnologías de la Web Semántica, estándares recomendados por W3C y basándonos en el enfoque propuesto por LOD.

El desarrollo realizado ha aportado resultados concretos al proyecto de investigación. En particular hemos cumplido con los siguientes hitos:

- La implementación de un mecanismo de cita a datos dentro del RDI.
- La evaluación de métodos de extracción, data-cleaning y publicación en RDF del RDI DSpace y del SNDB accesibles a través de SPARQL endpoint públicos.
- Un prototipo de infraestructura con un conjunto de herramientas bien definidas, que resulta adecuada para los desarrollos futuros.
- Definición de una consulta integrada SPARQL que involucra dos o más conjuntos de datos RDF y que resuelve la vinculación de publicaciones científicas con datos primarios citados en ella.

Los trabajos futuros, referidos a los temas desarrollados en este trabajo, se enfocarán al refinamiento de citas a datos, con vinculación bidireccional, de las publicaciones a los datos y de los datos a las publicaciones. Otro aspecto importante relacionado con la semántica será desarrollar ontologías propias del dominio que resulten necesarias para el RDI de la UNPSJB.

6. Referencias

Altman, M., & King, G. (2008). A Proposed Standard for the Scholarly Citation of Quantitative Data. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1081955

Arano, S., Martínez, G., Losada, M., Villegas, M., Casaldàliga, A., & Bel, N. (2011). La comunidad «Recursos y datos primarios» de la Universitat Pompeu Fabra: los repositorios institucionales como infraestructuras científicas: estudio de caso. *Revista Española de Documentación Científica*, *34*(3), 385–407.

Ball, A., & Duke, M. (2011). How to cite datasets and link to publications. *Digital Curation Centre*.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). THE SEMANTIC WEB. *Scientific American*, *284*(5), 34–43.

Buneman, P., Davidson, S., & Frew, J. (2016). Why data citation is a computational

problem. *Communications of the ACM*, *59*(9), 50–57.
Davidson, S. B., Buneman, P., Deutch, D., Milo, T., & Silvello, G. (2017). Data Citation: a Computational Challenge. *Proceedings of the ... ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, *2017*, 1–4.
Green, T. (2009). We need publishing standards for datasets and data tables. *Learned Publishing: Journal of the Association of Learned and Professional Society Publishers*, *22*(4), 325–327.
Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web Into a Global Data Space*. Morgan & Claypool Publishers.
Hyland, B., Atemezing, G., & Villazón-Terrazas, B. (2014). Best practices for publishing linked data. *W3C Working Group Note*.
Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman, I. I., & Others. (2014). Five stars of linked data vocabulary use. *Semantic Web*, *5*(3), 173–176.
Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. *International Journal of Digital Curation*, *6*(2), 4–37.
Remsen, K. D. M. R. T. (2011). D, Braak. *Darwin Core Archive How-To Guide*.
Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., … Desmet, P. (2014). The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PloS One*, *9*(8), e102623.
Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., … Walker, J. H. (2003). DSpace: An Open Source Dynamic Digital Repository. *DSpace: An Open Source Dynamic Digital Repository. Smith, MacKenzie; D-Lib Magazine*, *9*(1). https://doi.org/10.1045/january2003-smith
Starr, J., & Gastl, A. (2011). isCitedBy: A Metadata Scheme for DataCite. Retrieved from https://escholarship.org/uc/item/6r03h784
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PloS One*, *7*(1), e29715.
Zárate, M., Braun, G. A., & Fillottrani, P. R. (2017). Adding Biodiversity Datasets from Argentinian Patagonia to the Web of Data. In *International Semantic Web Conference (Posters, Demos & Industry Tracks)*. Retrieved from http://ceur-ws.org/Vol-1933/paper-6.pdf

Datos y Referencias Biográficas de los Autores

- ***Carlos Buckle***

  e-mail: *cbuckle@unpata.edu.ar* *y* *carlos.buckle@gmail.com*

  Antecedentes:

  Profesor Asociado en la carrera de Licenciatura en Informática de la Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB). Es docente responsable de las cátedras de Sistemas Operativos y Bases de Datos, en las cuales desarrolla su tarea de investigación. Es investigador categoría III en el programa de Incentivos. Actualmente dirige el proyecto "Infraestructura de Acceso a Datos Primarios con aporte de semántica en Repositorios Digitales" e integra el proyecto "Procesamiento y análisis de datos espaciales y temporales relativos a entornos urbanos".
  Cuenta con publicaciones en libros, revistas y congresos, producidas en diferentes proyectos de investigación en los que ha participado y participa.
  Es el representante de la UNPSJB ante la Red UNCI (Red de Universidades con Carreras de Informática de Argentina)
  Es coordinador de workshops, integrante del Comité Académico-Científico y Evaluador de Trabajos en congresos de la disciplina. Asimismo ha participado en la evaluación de recursos humanos como jurado en concursos docentes en universidades argentinas.
  En formación de recursos, ha participado como tutor de Tesinas de grado y Pasantías.
  En Extensión Universitaria, es Director del proyecto "Integración Nivel Medio Madryn" financiado por la Secretaría de Políticas Universitarias y ha participado como director e integrante en proyectos de extensión, vinculación y transferencia tecnológica anteriormente, en forma sostenida.
  En Gestión Universitaria integra actualmente el Consejo Zonal de la Sede Puerto Madryn y es consiliario suplente en el Consejo Superior de la UNPSJB. Ha ocupado cargos en cuerpos colegiados de la institución y cargos ejecutivos como Delegado Rectoral en Puerto Madryn y Responsable del Departamento de Informática.
  En ámbitos no-universitarios, se ha desempeñado como responsable de equipo de desarrollo en empresas, como director de centros de cómputo en organismos del estado y como diseñador/desarrollador de soluciones informáticas en diferentes organismos públicos y privados.

- ***Renato Mazzanti***

  e-mail: *renato@cenpat-conicet.gob.ar* *y* *renato.mazzanti@gmail.com*

  Antecedentes:

  Profesor Adjunto de las cátedras: "Programación Orientada a Objetos" y "Algorítmica y Programación II". Es alumno de la Especialización en Gestión de la Información Científica y Tecnológica" de la Facultad de Humanidades

y Ciencias de la Educación de la Universidad Nacional de la Plata. Asesor informático de los Consejos Asesores: “Sistema Nacional de Datos Biológicos (SNDB)” y “Sistema Nacional de Datos de Mar (SNDM)” del Ministerio de Ciencia, Tecnología e Innovación Productiva. Coordinador de la Unidad de Gestión de la Información CCT CONICET-CENPAT.
Cuenta con publicaciones en revistas, congresos y workshops producidas en diferentes proyectos de investigación en los que ha participado y participa.
En formación de recursos humanos, ha participado y participa como tutor de Tesinas de grado y Pasantías.
En Gestión Universitaria integró el Consejo Zonal de la Sede Puerto Madryn y de la UNPSJB. En el pasado ha ocupado cargos en cuerpos colegiados de la institución y cargos académicos como Responsable del Departamento de Informática en dicha ciudad.
En ámbitos no-universitarios, se ha desempeñado como desarrollador en organismos privados y como Jefe del Servicio Centralizado de Cómputo en el CCT CONICET-CENPAT.

- ***Marcos Zárate***
  e-mail: *zarate@cenpat-conicet.gob.ar*
  El Lic. Marcos Zárate es alumno del Doctorado en Ciencias de la Computación de la Universidad Nacional del Sur y becario doctoral CONICET dentro del Centro Nacional Patagónico. Profesor auxiliar en la carrera Licenciatura en Informática de la UNPSJB sede Puerto Madryn, en la cátedra Algorítmica y Programación 2. Actualmente su área de investigación se centra en tecnologías asociadas a la Web Semántica y datos enlazados, se encuentra trabajando en colaboración con el Grupo de Investigación en Lenguajes e Inteligencia Artificial (GILIA) de la Universidad Nacional del Comahue y con el Departamento de Ciencias e Ingeniería de la Computación (DCIC) de la Universidad Nacional del Sur. Se desempeña como experto informático designado por resolución 044/16 en el Sistema Nacional de Datos Biológicos del MINCyT.

- ***Gustavo Samec***
  e-mail: *gsamec@cenpat-conicet.gob.ar* *y* *gsamec@gmail.com*
  Antecedentes:
  Jefe de Trabajos Prácticos en la carrera de Licenciatura en Informática de la UNPSJB, en las cátedras de Algorítmica y Programación II y Programación Orientada a Objetos.
  Cuenta con publicaciones en congresos y workshops producidas en diferentes proyectos de investigación en los que ha participado y participa.

En formación de recursos humanos, ha participado y participa como tutor de pasantías.
Actualmente realiza desarrollos e investigación en Repositorios de Datos, su integración con otras aplicaciones y el manejo de grandes volúmenes de datos en los mismos.
En ámbitos no-universitarios, desempeña funciones como Personal de Apoyo en el CCT CONICET-CENPAT donde está a cargo del Servicio Centralizado de Computación.