Conectando fortalezas y capacidades institucionales: automatización de la integración VIVO-Pure-DSpace

MALGORZATA LISOWSKA NAVARRO
Universidad del Rosario (Colombia)
margarita.lisowska@urosario.edu.co

HUMBERTO BLANCO CASTILLO
Universidad del Rosario (Colombia)
humberto.blanco@urosario.edu.co

ÁNGELICA CHARTANO HERNÁNDEZ
Universidad del Rosario (Colombia)
angelica.chartano@urosario.edu.co

JUAN FELIPE LÓPEZ
Universidad del Rosario (Colombia)
juan.lopez@urosario.edu.co

RESUMEN

El HUB-UR: Services and Experts Finder, portal de fortalezas y capacidades de la Universidad del Rosario, implementado en el año 2018 por el Centro de Recursos para el Aprendizaje y la Investigación (CRAI), recibe anualmente más de 110.000 visitas de cien países. El HUB-UR, basado en el software de web semántica y de código abierto VIVO permite, a través de búsquedas y filtros, descubrir las fortalezas y capacidades en investigación, docencia y extensión de la universidad, aumentando así su reconocimiento a nivel nacional e internacional. En

sus tres años de funcionamiento, el reto continúa siendo obtener la información actualizada y automatizada de otras fuentes institucionales como el Repositorio Institucional E-docUR (DSpace) y el sistema Pure, software de la casa editorial Elsevier, a través del cual se gestiona la investigación en la universidad. Dado que no existe abundante literatura acerca de la interoperabilidad de VIVO con estas plataformas, la presente ponencia pretende ilustrar la forma a través de la cual se ha desarrollado la integración entre VIVO, Pure y DSpace. Nuestro interés se centra en compartir la experiencia y los retos que se presentaron en el proceso de automatización y que puedan servir a otras instituciones interesadas en este proceso de integración.

PALABRAS CLAVE

Interoperabilidad; repositorios académicos; visibilidad académica; VIVO.

Interoperability; academic repositories; academic visibility; DSpace; Pure.

El HUB-UR: Services and Experts Finder, basado en el software VIVO aprovecha las ventajas de datos enlazados abiertos (Linked Open Data)[1], para relacionar y vincular la información de perfiles individuales, institucionales, proyectos, productos de investigación y eventos, entre otros (CONFLUENCE, 2022), proporcionando así resultados hipervinculados en forma de servicios para diferentes públicos objetivo. Por ejemplo, información para estudiantes en proceso de grado a través del servicio "*Find a supervisor*" o información para pares de investigación a través del servicio "*Find a research partner*" .

[1] Linked Data se refiere a datos enlazados a través de web semántica, la manera de describir y representar esta información se realiza mediante el estándar RDF (Resource Description Framework).

Para visibilizar las capacidades y fortalezas tanto individuales como institucionales, fue necesario integrar diferentes fuentes de información institucional (ver GRÁFICO 1), esto con el fin de asegurar su calidad, actualización y confiabilidad.

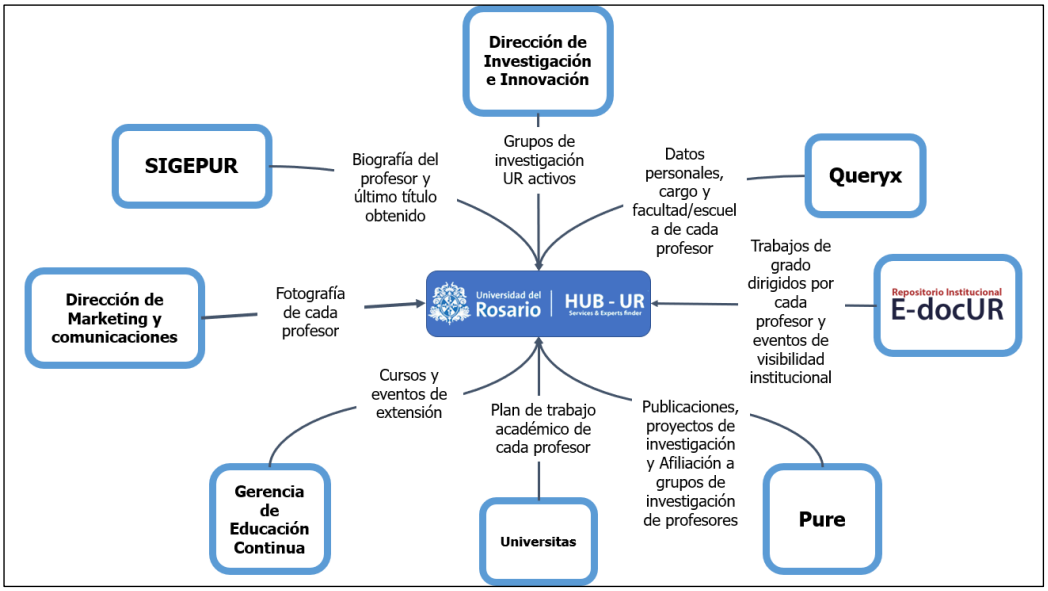

GRÁFICO 1. Diagrama de sistemas. Fuente: HUB-UR

Esta integración ha constituido un reto, pues los sistemas de información fuente han sido desarrollados en diferentes momentos de la historia de la universidad, lo que conlleva al registro de información de diferentes maneras; por esta razón, algunos no cuentan con interfaces que permitan la interoperabilidad, sino que, de forma manual, se realiza la extracción, transformación y carga de información en VIVO.

Actualmente, se ha logrado la automatización con los sistemas de gestión de investigación Pure de Elsevier y DSpace, los cuales proveen principalmente información sobre producción académica y actividad científica de la universidad. El Sistema de Gestión de la Investigación - Pure[2] es la plataforma

[2] Pure: https://www.elsevier.com/solutions/pure

utilizada para la gestión de la investigación en la UR, que contiene información sobre proyectos y resultados de investigación; actualmente el software Pure se encuentra en la versión 5.21. Por otra parte, el Repositorio Institucional E-docUR, es la plataforma institucional que almacena, preserva y difunde la producción institucional en su función docente, investigativa y de extensión, está implementado en el software de código abierto DSpace en su versión 6.3.

Proceso de automatización de la integración VIVO-DSpace-Pure

1. Integración del sistema de gestión de investigación UR-Pure-VIVO

El proceso de extracción y carga de información en Pure se ha realizado de acuerdo con el siguiente diagrama (DIAGRAMA 1).

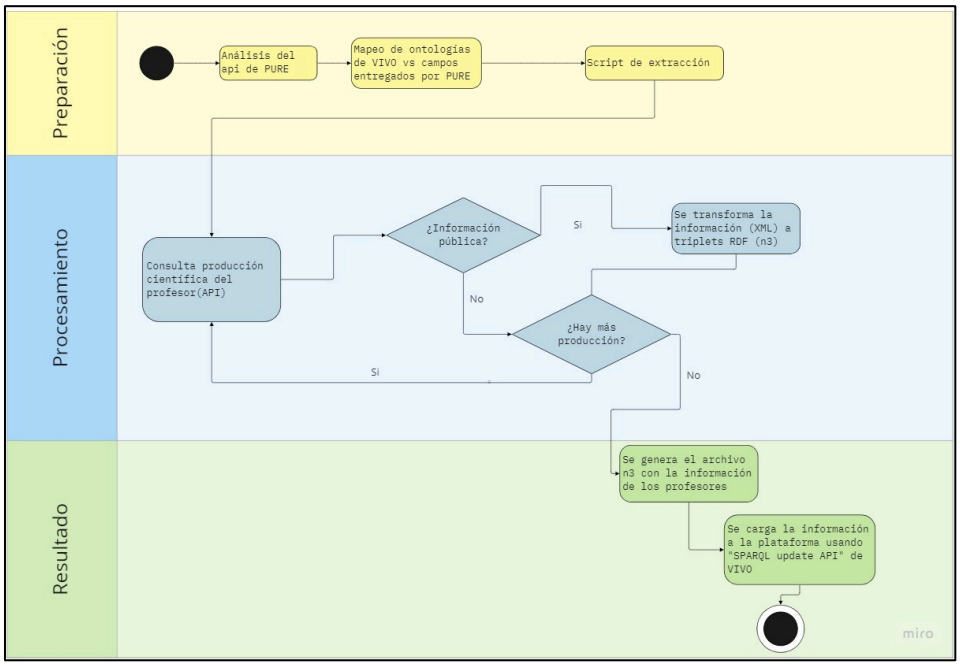

DIAGRAMA 1. Etapas de integración VIVO-Pure

La fase de preparación para implementar la solución partió del análisis de la API de Pure (ELSEVIER, 2022), para ello fue necesario solicitar al proveedor del servicio (Elsevier) una API KEY, realizar su conexión y empezar la exploración de la API. En esta fase, fue necesario realizar ejercicios para validar la forma y contenidos en que la API entregaba la información para ser consumidos, además de los requerimientos para su conexión.

Posteriormente, fue necesario realizar un mapeo de la información disponible en Pure para determinar su correspondiente ontología en VIVO, requerida para almacenar la información que sería extraída. A continuación, se presenta un ejemplo del mapeo realizado (TABLA 1):

ATRIBUTO	ONTOLOGÍA
Titulo	*22-rdf-schema:label*
Keywords	*vivo:freetextkeyword*
Tipo	*22-rdf-syntax-ns:type*
Número de páginas	*bibo:numpages*
Abstract	*Core:overview*
autores	*Core:relatedBy*
Volumen	*bibo:volume*
Edición	*bibo:issue*
ISSN	*bibo:issn*
DOI	*Vcard:hasURL*
Idioma	*vcard:language*
Journal	*vlocal:academicarticlejournal*
Fecha de publicación	*vivo:datetimevalue*
Página de inicio	*bibo:pagestart*

ATRIBUTO	ONTOLOGÍA
ISBN	*vlocal:isbn*
Evento	*vlocal:conferencePaperEvent*
Número de patente	*core:patentNumber*

TABLA 1. Mapeo de atributos Pure-Ontología VIVO

La fase que contempla la transformación de los atributos a triples al inicio del proyecto, fue realizada empleando el software Karma, el cual permite generar los archivos con la información procesada para ser cargada a VIVO (CONLON, 2021), sin embargo, este proceso fue optimizado posteriormente con la creación de un *script* que automatizó el proceso de extracción de información del API de PURE y su transformación en tripletas RDF, formato n3 (W3C, 2011).

La fase de procesamiento que inicia desde la extracción con la consulta a la API de Pure de los datos personales e identifica de la producción de cada uno de los de los profesores activos en el HUB-UR, esta consulta se realiza con base en el documento de identificación (ID) de cada profesor.

El *script* se encarga de mapear la respuesta entregada por el API de Pure a las ontologías correspondientes de VIVO, esto garantiza que los productos de investigación y sus metadatos tengan las clases y propiedades de datos correspondientes en el HUB-UR. Al finalizar este procesamiento su resultado es un archivo generado por el *script* en formato n3 con la información recopilada de los productos, los cuales posteriormente son cargados a VIVO utilizando el "SPARQL update API", que brinda capacidades para consultar patrones gráficos requeridos y opcionales junto con sus conjunciones y disyunciones (W3C, 2022).

2. Integración del repositorio institucional E-docUR (DSpace)-VIVO

Como se ha indicado, el Repositorio institucional E-docUR (DSpace) alberga dentro de sus contenidos los trabajos de grado y tesis que han sido dirigidas por los profesores UR así como los productos asociados a su participación en conferencias y eventos, estos contenidos al considerarlos parte de las fortalezas y capacidades de los profesores son integrados en el HUB-UR.

La extracción de la información que proviene de DSpace es realizada directamente desde la base de datos del repositorio, y se extrae por medio de *scripts* de acuerdo con el contenido a extraer; uno para las tesis y trabajos de grado dirigidos, otro para las conferencias y otro para los eventos.

Para la extracción de tesis y trabajos de grado la consulta es realizada a través del documento de identidad del profesor (ID) que para DSpace corresponde al valor "Authority" de la tabla *metadatavalue*. El siguiente diagrama muestra las diferentes etapas en el proceso (DIAGRAMA 2).

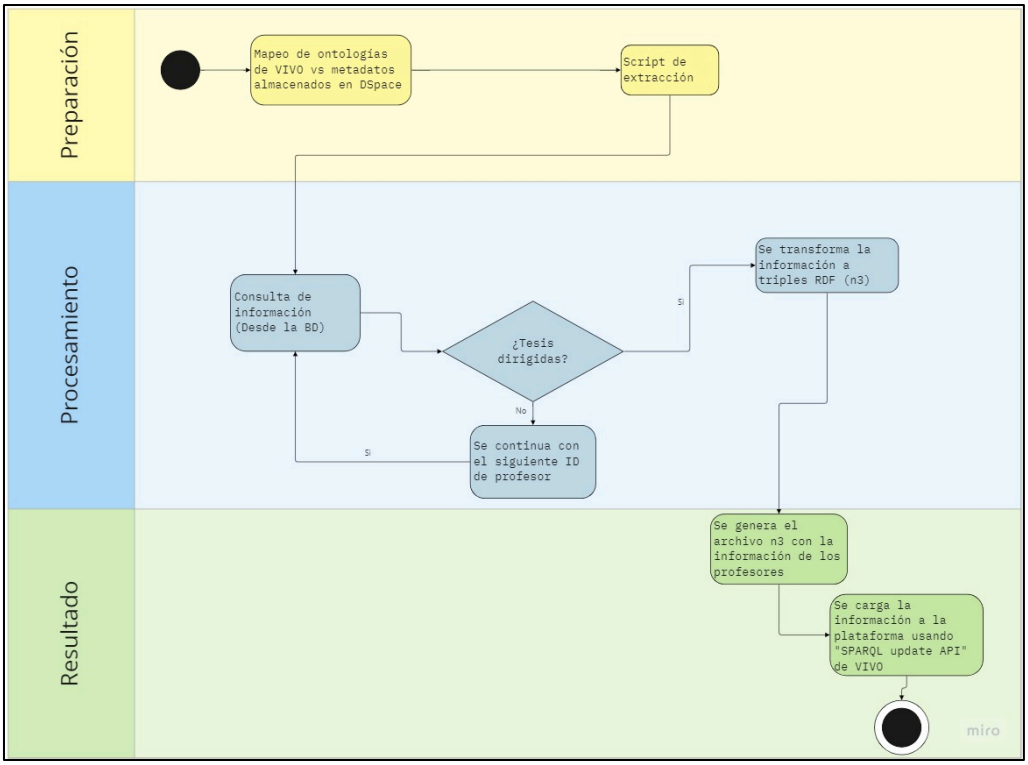

DIAGRAMA 2. Etapas de integración VIVO-DSpace (tesis y trabajos de grado)

Para el caso de las tesis y trabajos de grado el mapeo de los atributos en DSpace y su correspondiente ontología en VIVO se realiza sobre los siguientes campos (TABLA 2):

ATRIBUTO	ONTOLOGÍA
Titulo	*22-rdf-schema:label*
Keywords	*vivo:freetextkeyword*
Tipo	*22-rdf-syntax-ns:type*
Abstract	*Core:overview*
autores	*Core:relatedBy*
Idioma	*vcard:language*
Fecha de publicación	*vivo:datetimevalue*

TABLA 2. Ejemplo de mapeo atributo DSpace -Ontología VIVO (tesis y trabajos de grado)

Una vez consolidada y procesada la información se envía a través de la API de actualización de VIVO donde el resultado es el archivo generado por el *script* en formato n3, el cual es cargado a VIVO utilizando el "SPARQL update API".

Para la extracción de eventos y conferencias es necesario realizar una identificación previa de las colecciones que alojan esos productos en DSpace, a partir de esta identificación el *script* realiza la consulta de los participantes (autores) en los eventos y conferencias y valida su existencia en un perfil de VIVO. Al ser identificado, el *script* genera el triples con los datos del evento o conferencia (DIAGRAMA 3).

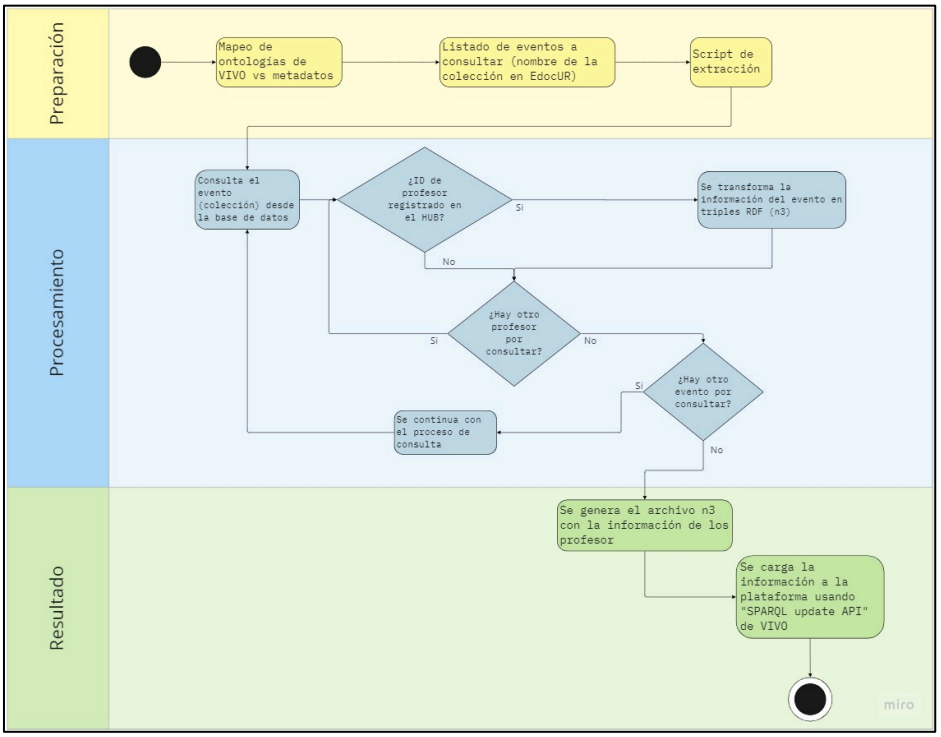

DIAGRAMA 3. Etapas de integración VIVO-DSpace (eventos y conferencia)

Para los eventos y conferencias los atributos en DSPACE y su correspondiente ontología en VIVO es la siguiente tabla:

ATRIBUTO	ONTOLOGÍA
Titulo	*22-rdf-schema:label*
Keywords	*vivo:freetextkeyword*
Tipo	*22-rdf-syntax-ns:type*
Abstract	*Core:overview*
Temáticas	*vivo:hassubjectarea*
Participantes	*vlocal:participants*
Descripción	*vivo:description*
Fecha	*vivo:datetimeinterval*

TABLA 3. Ejemplo de mapeo atributo DSpace -Ontología VIVO (eventos y conferencias)

Retos identificados en el proceso

El proceso de extracción carga y actualización de la información ha supuesto retos que se han superado gracias a los aprendizajes obtenidos a medida que el proyecto ha ido madurando. Algunos de ellos se presentan a continuación:

- Traducir la información de Pure y DSpace, a un lenguaje comprensible para VIVO fue un proceso complejo. Esto debido a que este último utiliza ontologías que requieren archivos en formato n3. Para lograr esto ha sido necesario comprender las consultas SPARQL, con el fin de identificar las clases, relaciones y ontologías que se deben asignar a cada tipología extraída de Pure y DSpace.
- Para garantizar la funcionalidad del procedimiento ha sido necesario estar alineados a los cambios generados en las nuevas versiones de la API de Pure, lo que implica un monitoreo permanente y aplicar cambios periódicos en los *scripts*.
- El trabajo interdisciplinar con otras áreas exige una comunicación constante y el diseño de flujos de trabajo que permita dar respuesta oportuna ante los cambios en las plataformas.

Impactos obtenidos

La implementación de estos procedimientos generó resultados significativos, dentro de los que se encuentran:

- Reducción en los tiempos de gestión: al automatizar los procesos de extracción de información, se logró reducir los tiempos de gestión que implicaban revisiones y ajustes manuales de la información, especialmente en el momento de actualización de contenidos.
- Información actualizada y fiable: durante el proceso de extracción de información de sistemas fuente, ha permitido identificar información incompleta, duplicada o desactualizada, que ha implicado el trabajo conjunto de depuración y corrección de la información con las áreas responsables con lo cual se ha logrado unificar la información

institucional y mejorar la calidad de la información disponible en los sistemas fuente.

- Reducción en frecuencia de actualización de productos: al automatizarse la extracción de información a través de los *scripts* la actualización de productos en VIVO pasó de ser semestral a trimestral lo que permite una mayor correspondencia en los productos asociados a cada profesor en las diferentes plataformas.

Se evidencia mayor concordancia institucional sobre la información visible relacionada con la producción académica y de investigación.

Mejoras y trabajo futuro

Algunas de las mejoras que se han contemplado desde el aspecto tecnológico para el HUB-UR: Services and Experts Finder son:

- Actualmente se trabaja en la mejora de los *scripts* para que la sincronización de la información proveniente de Pure y DSpace se realice de forma inmediata.
- Se ha evaluado la creación de un nuevo *script* que permita identificar y eliminar de forma automática los registros que dejan de existir en Pure, lo cual obedece a procesos como depuraciones, correcciones o a la salida de profesores de la planta.
- Se evalúa la integración con nuevas fuentes de información institucionales que pueden proveer contenidos para VIVO.
- La interoperabilidad entre las plataformas VIVO-DSpace-Pure, permitirá en un futuro la visualización de objetos digitales complejos en el HUB-UR que faciliten la identificación y vinculación de diferentes productos vinculados al mismo proceso de investigación.

Bibliografía

CONFLUENCE. (2022). VIVO. VIVO 1.12.x. https://wiki.lyrasis.org/display/VIVODOC112x/SPARQL+Update+API

CONLON, M. (2021). Karma for Data Ingestion—VIVO - LYRASIS Wiki [Wiki]. Confluence. https://wiki.lyrasis.org/display/VIVO/Karma+for+Data+Ingestion

ELSEVIER. (2022). PURE API - 5.21. Pure API. https://purehost.bath.ac.uk/ws/api/521/api-docs/documentation/Content/Topics/Web_Services_Intro.htm

W3C. (2011). Notation3 (N3): A readable RDF syntax https://www.w3.org/TeamSubmission/n3/

W3C. (2022). SPARQL Query Language for RDF. https://www.w3.org/TR/rdf-sparql-query/