Vol. 20, n.° 3, pp. 155-166, 2021
Revista UIS Ingenierías
Página de la revista: https://revistas.uis.edu.co/index.php/revistauisingenierias

Databio, galería de biodiversidad del Jardín Botánico Juan María Céspedes

Databio, biodiversity gallery of the Juan María Céspedes Botanical Garden

Danny Alexander Carrillo-Torres [1a], Angelica Ocampo-Escobar [1b], Royer David Estrada-Esponda [1c]

[1] Escuela de Ingeniería de sistemas y Computación, Universidad del Valle, sede Tuluá, Colombia.
Correos electrónicos: [a] danny.carrillo@correounivalle.edu.co, [b] ocampo.angelica@correounivalle.edu.co,
[c] royer.estrada@correounivalle.edu.co. Orcid: [a] 0000-0003-2376-7979, [b] 0000-0002-7321-7468,
[c] 0000-0002-6849-1278

Recibido: 3 octubre, 2020. Aceptado: 12 marzo, 2021. Versión final: 2 junio, 2021.

Resumen

La biodiversidad es esencial para la vida y motiva muchos esfuerzos para su conservación, uno de esos esfuerzos es la divulgación de datos sobre especímenes vegetales, que son utilizados por instituciones como el Instituto para la Investigación y Preservación del Patrimonio Cultural y Natural del Valle del Cauca (INCIVA). Aunque este último desarrolla, estimula y apoya procesos de apropiación, generación y divulgación del conocimiento para la conservación y uso del patrimonio natural y cultural del Valle del Cauca, no soporta esos procesos con herramientas tecnológicas adecuadas que, integrando técnicas especializadas, permitan la divulgación del conocimiento como lo promueve la iniciativa Linked Open Data (LOD). Esta permite la publicación y el enlazamiento de información en la Web de Datos. La principal contribución de este artículo es la presentación de DATABIO, una aplicación web que administra, publica y divulga información de la colección biológica del jardín botánico por medio de un Dataset enlazado y publicado en la web de datos que como característica principal facilita la actualización de sus tripletas cada vez que hay nuevos registros en el sistema transaccional. DATABIO fue desarrollado gracias al uso de programación extrema como metodología de desarrollo de software, pero previamente se incluyeron fases para el modelamiento, almacenamiento, enlazamiento y publicación de datos. Como resultado principal se destaca que la información publicada cumple con el esquema de datos abierto de Tim Berners-Lee obteniendo cinco estrellas, lo cual facilita la toma de decisiones a instituciones que promueven la conservación del medio ambiente.

Palabras clave: colecciones biológicas; Linked Open Data; Resource Description Framework; sistematización; SPARQL.

Abstract

The biodiversity is essential for life and it motivates many efforts for its conservation. One of those efforts is the dissemination of data on plant specimens, which institutions such as the Institute for Research and Preservation of the Cultural and Natural Heritage of the Valle del Cauca (INCIVA) use. The institute develops, stimulates, and supports processes of appropriation, generation, and dissemination of knowledge for the conservation as well as the use of the natural and cultural heritage of Valle del Cauca. However, it does not support these processes with technological tools that integrate specialized techniques that allow the dissemination of knowledge such as the Linked Open Data (LOD). This one allows the publication and linking of information on the Data Web. Therefore, the article presents DATABIO. It is a web application that manages, publishes, and disseminates information from the biological collection of the botanical garden. It is done using a Dataset linked and published on the data web that as the main feature facilitates the updating of its triples every time there are new records in the transactional system. DATABIO was developed

ISSN impreso: 1657 - 4583. ISSN en línea: 2145 - 8456,
Como citar: D. A. Carrillo-Torres, A. Ocampo-Escobar, R. D. Estrada-Esponda, “Databio, galería de biodiversidad del Jardín Botánico Juan María Céspedes,” *Rev. UIS Ing.*, vol. 20, no. 3, pp. 155-166, 2021, doi: 10.18273/revuin.v20n3-2021011

thanks to the use of extreme programming as a software development methodology, but it previously required the application of phases for modeling, storage, linking, and data publishing. As a main result, it stands out that the information published complies with the open data scheme of Tim Berners -Lee, obtaining five stars, which facilitates decision-making for institutions that promote environmental conservation.

Keywords: biological collections; Linked Open Data; Resource Description Framework; systematization; SPARQL.

1. Introducción

Las colecciones biológicas son un conjunto de especímenes, o partes de éstos, organizados con el fin de proporcionar información sobre la procedencia, la colecta e identificación de cada uno de ellos y cumplen un papel fundamental para la conservación del patrimonio biológico [1], [2]. Por otro lado, la administración, publicación y divulgación de estos conjuntos de datos en plataformas se conoce como sistematización de colecciones biológicas [3] y tienen como objetivo la consulta y la preservación de dichas colecciones a través de diferentes tecnologías, por ejemplo, una herramienta online de publicación IPT (Integrated Publishing Toolkit) [4] que permite incorporar y compartir conjuntos de datos de biodiversidad compatible a través de internet.

En la actualidad hay instituciones en Colombia que tienen colecciones biológicas asociadas a especímenes vegetales [5]. El Herbario de la Universidad del Quindío es un ejemplo de ello, allí desarrollaron una plataforma web compuesta por módulos administrativos, de información y museo virtual., También consideraron procesos de obtención de información, contrastación de datos, conservación de especímenes, intercambio entre expertos, difusión con la comunidad y de investigación y modelamiento [6]. Otras plataformas también exponen utilidades funcionales importantes para favorecer la conservación de diferentes tipos de colecciones, por ejemplo, el modelamiento digital de especímenes [7] y exportación en diferentes formatos no propietarios como csv y estándares como Darwin Core [8].

Por otor lado, el Instituto para la Investigación y la Preservación del Patrimonio Cultural y Natural del Valle del Cauca (INCIVA), en su visión expone entre otras cosas, ser reconocida regional, nacional e internacionalmente por la generación y divulgación del conocimiento [9]. Para esto puede resultar pertinente crear procesos y mecanismos de gestión que aceleren los procesos de aprendizaje, la creación, la adaptación y la difusión del conocimiento. Estas tareas están recogidas en lo que se conoce como gestión del conocimiento, que se basa en el reconocimiento y la utilización del valor más importante de las organizaciones: los recursos humanos, el conocimiento y la disposición de colocarlo al servicio de la comunidad [10].

Para una adecuada gestión del conocimiento existen procesos estratégicos, dentro de los cuales está la divulgación del conocimiento. Este proceso estratégico es fundamental para la contribución de la misión organizacional de cualquier institución [10]. Si bien el conocimiento se puede divulgar a través de acciones personales, estas se pueden apoyar o soportar en herramientas tecnológicas que facilitan compartir o divulgar el conocimiento.

Actualmente, el Jardín Botánico Juan María Céspedes de Tuluá, utiliza herramientas tecnológicas dispersas para contribuir con dicho proceso estratégico. La figura 1 relaciona esas herramientas.

Figura 1. Herramientas utilizadas actualmente por el Jardín Botánico.

Teniendo en cuenta que la administración de forma dispersa de la información ralentiza los procesos misionales del Jardín y va en contravía de nuevas tendencias que potencian la apropiación social del conocimiento, es necesario identificar dichas herramientas e integrarlas para potenciar los resultados de la organización en términos de divulgación y lo que ello implica. De hecho, dicha integración mejoraría la experiencia de los usuarios ya que es sustancialmente mejor tener integrados varios servicios que utilizar herramientas de forma dispersa [11]. Una de estas tendencias es la Web Semántica. Esta tecnología es útil para la divulgación de cualquier tipo de información, ya que su objetivo es dotar a las aplicaciones informáticas

de la capacidad de estructurar y manejar la información con base a una valoración semántica de sus contenidos. Esto proporciona la capacidad de entender el contenido, posibilitando que las personas trabajen más eficientemente en un marco cooperativo [12].

Con el fin de lograr ese objetivo, es necesario modificar la forma en la que trabaja la Web, perfeccionando y ampliando sus capacidades semánticas. Para ello existen diversos lenguajes que estructuran la información de manera similar al análisis semántico del lenguaje humano, por ejemplo, el lenguaje Resource Description Framework (RDF). RDF es un marco de trabajo en el que es posible describir información para que la procesen las máquinas e intentar aportar significado a la estructura de los documentos [13]. Este último es un componente del Linked Open Data (LOD) que es un método de publicación de datos estructurados para que puedan ser interconectados y más útiles. Este método permite que sean conectados y consultados datos de diferentes fuentes [14].

Otro elemento clave de la Web Semántica son las ontologías, descritas como sistemas de expresiones que definen formalmente las relaciones entre términos. En este contexto, el objetivo fundamental de las ontologías es conseguir la desambiguación del sentido de las palabras en un documento, es decir, identificar el significado correcto de las palabras polisémicas según el contexto de los documentos [15]. Para el caso del Jardín Botánico y otras instituciones dedicadas a preservar colecciones biológicas o incluso otro tipo de colecciones, existe el formato Darwin Core, que es una ontología para facilitar la caracterización, procesamiento e interpretación de los individuos o especies que hacen parte de dichas colecciones. Una explicación amplia sobre dicho estándar se encuentra en [16].

En el escenario de esas tecnologías, el Gobierno Colombiano dispone de plataformas para la publicación de la información en el marco de una estrategia llamada Gobierno Abierto, que consiste en publicar Datos de acceso Abierto disponibles para toda la comunidad en general [17], es decir, datos publicados en formatos digitales estandarizados. Esto es valioso en términos de apropiación social de conocimiento y la contribución con la reducción de asimetrías sociales. Sin embargo, en términos especializados se encuentra que no hay iniciativas que involucren el uso de la LOD para los datos referentes a colecciones de especímenes vegetales, lo cual sumado a la ya mencionada administración de la información que realiza de manera dispersa el Jardín Botánico Juan María Céspedes se traduce en una oportunidad para integrar no solo los procesos de sistematización de colecciones vegetales, sino que también contribuir con su divulgación y a la vez a con la generación de conocimiento no solo especializado sino que también de dominio general para promover el cuidado del medio ambiente y su Biodiversidad.

Es así como a continuación se presenta Databio, una aplicación Web que además de favorecer la sistematización de colecciones facilita su divulgación por medio de un Dataset publicado y divulgado en la Web de datos. Es decir, el objetivo principal del proyecto consistió en elaborar un Dataset con información referente a los especímenes vegetales del Jardín Botánico del Valle Juan María Céspedes en el municipio de Tuluá, bajo los principios del LOD con el ánimo de minimizar el problema de acceso a la información manifestado en [2] no solo para comunidades científicas, sino que también para la comunidad en general y que claramente va en contravía de la difusión del conocimiento sobre biodiversidad en el país.

En términos del contexto geográfico, el Jardín Botánico está localizado en el corregimiento de Mateguadua, a 7 Km de la cabecera municipal de Tuluá. Tiene una extensión de 154 hectáreas de terrenos ondulados, con alturas que oscilan entre los 1.050 y 1.300 metros sobre el nivel del mar. La temperatura promedio es la misma de la planicie central del Valle del Cauca, de unos 25 °C y la precipitación pluvial promedia unos 1.000 mm. anuales, con dos períodos trimestrales de lluvias alternados con dos de sequía. Está ubicado dentro de la zona de vida denominada Bosque Seco Tropical (B.s.t) [18].

Finalmente, el artículo continuo con la presentación de la metodología de investigación, así como la metodología de publicación y enlazamiento en la web de datos. Después presenta los resultados en el contexto de la última metodología mencionada. Posteriormente se presentan algunos aspectos para discusión, así como las conclusiones y trabajos futuros. Por último, se presentan las referencias bibliográficas.

2. Metodología

Teniendo en cuenta que el desarrollo de la aplicación web incluyó elementos relacionados con LOD, fue necesario trabajar desde diferentes frameworks, el primero asociado al desarrollo de sistemas informáticos y el segundo en el ámbito de la publicación de datos, ambos en el contexto de una investigación aplicada de base tecnológica con enfoques exploratorios y descriptivos. La figura 2 detalla el proceso metodológico propuesto en [19].

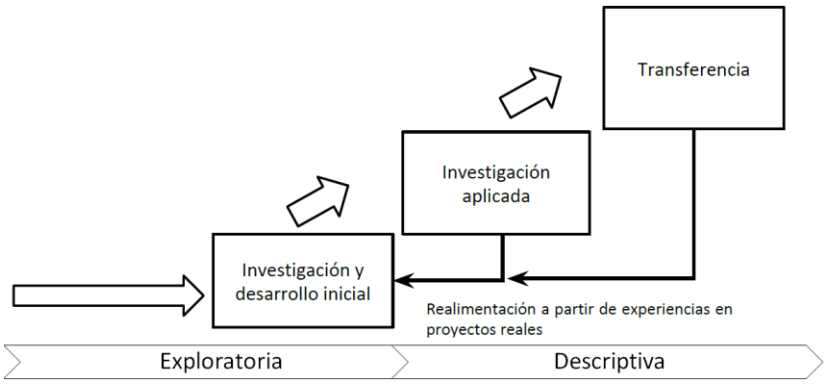

Figura 2. Metodología.

En cuanto a la investigación y desarrollo inicial se participó en reuniones con personal del Jardín Botánico del Valle Juan María Céspedes, esto con el fin de establecer el alcance del proyecto y definir necesidades de los diferentes interesados del mismo. Adicionalmente, se revisaron otros trabajos que sirvieron como antecedentes para la formulación y posterior ejecución del proyecto de Investigación y Desarrollo.

Por otro lado, la investigación aplicada incluyó toda la validación de componentes y tecnologías que permitieron cumplir con el objetivo principal del proyecto, es así que durante esta fase se revisaron lenguajes de programación, aplicaciones de terceros para publicación y enlazado de datos, textos académicos, marcos de trabajo y demás actividades que condujeron a aumentos de valor incrementales para los diferentes interesados del proyecto.

En términos de transferencia fue posible realizar divulgaciones parciales en algunos eventos académicos tales como la semana de expresión Universitaria de la Universidad del Valle sede Tuluá, el Encuentro Departamental de Semilleros de Investigación, evento que se desarrolló el mes de mayo de 2019 en el ITA profesional de la ciudad de Buga y el noveno encuentro Regional de Semilleros de Investigación, RREDSI en octubre del 2019 realizado en Tuluá, Valle del Cauca. Estas actividades sirvieron para enriquecer el trabajo que aquí se presenta.

De manera transversal, desde el enfoque exploratorio se buscó clarificar las ideas que posibilitaron responder a las preguntas de investigación y materializar los insumos necesarios para plantear sus soluciones; adicionalmente, se abordó un segundo enfoque transversal de naturaleza descriptivo, el cual consistió en describir los resultados producto del ejercicio de exploración por medio de la recopilación y tabulación de sus resultados. Como se aprecia en la figura 2. Las fases se repitieron según las necesidades que fueron presentando y los enfoques exploratorios y descriptivos fueron la plataforma para la ejecución de dichas etapas.

En cuanto a la metodología de Desarrollo de Software se postularon cuatro de ellas, todas ágiles, las cuales fueron: Scrum, XP, TDD e ICONIX. Se desarrolló una revisión de las cuatro metodologías que sirvió para poner en marcha una actividad que se centró en comparar las metodologías y asignarle una puntuación a cada una de ellas según unos criterios definidos.

Después de obtener los resultados de dicha actividad, clasificaron dos metodologías opcionadas para ser seleccionadas, las cuales fueron Scrum y Extreme Programming (XP), ambas con unos criterios que podían acoplarse fácilmente al presente proyecto, sin embargo, finalmente se optó por elegir la metodología de desarrollo de software Extreme Programming, ya que además de tener un framework adaptable, tiene unos objetivos claros, el ciclo de vida es liviano y por fases y su facilidad de uso es alta.

Como parte de la investigación aplicada se trabajó con un conjunto de datos estructurado en un formato Darwin Core proporcionado por el personal del INCIVA. Dicho conjunto al momento de la entrega contenía 4183 especímenes preservados en el Jardín Botánico Juan María Céspedes y que pertenecen a la colección biológica HERBARIO TULV.

La metodología que se siguió para el desarrollo del Dataset contempló cuatro fases que fueron formuladas gracias a la investigación encontrada en [20]. La figura 3 presenta dichas fases.

Figura 3. Fases metodológicas LOD.

3. Resultados

En las siguientes secciones se introducen los resultados más importantes de la ejecución de las fases metodológicas enunciadas.

3.1. Preprocesamiento de datos

Aunque los datos administrados por el Jardín se sistematizan en hojas de cálculo de paquetes ofimáticos bajo el estándar Darwin Core, fue necesario aplicar una fase de preprocesamiento. Esto es así porque no necesariamente todos los recolectores de especímenes diligencian la totalidad de campos propuestos por dicho estándar.

Para esta etapa se utilizó el lenguaje de programación Python y se importó Pandas que es una librería destinada al análisis de datos y proporciona unas estructuras de datos flexibles [21], que permiten trabajar con ellos de forma muy eficiente, todo esto con el fin de poder trabajar con datos limpios y listos para ser usados tanto en la creación del Dataset, como en la implementación de la aplicación.

Inicialmente, el archivo Excel que contiene la información sobre especímenes vegetales del Jardín Botánico con el estándar Darwin Core se convirtió a la extensión csv, esto para facilitar su preprocesamiento en Python con la librería mencionada anteriormente. Posterior a esto se convirtió la información almacenada en el archivo csv a un Data Frame, que es una estructura de datos perteneciente a la librería Pandas, similares a las tablas de bases de datos relacionales como SQL.

Teniendo la información organizada en un DataFrame, se agregan las columnas o atributos del estándar Darwin Core que hacían falta inicialmente en el Excel proporcionado inicialmente, puesto que es imprescindible contar con todos los atributos, ya que todos estos hacen parte de la Base de Datos diseñada e implementada en la aplicación. En el siguiente fragmento de código se muestra la inserción de las columnas.

```
import pandas as pd
data = pd.read_csv('dataset.csv')
df = pd.DataFrame(data)

nulo = ''

df.insert(6, 'type', nulo)
df.insert(7, 'modified', nulo)
df.insert(8, 'license', nulo)
df.insert(9, 'rightsHolder', nulo)
df.insert(10, 'accessRights', nulo)
df.insert(11, 'bibliographicCitation', nulo)
df.insert(12, 'references', nulo)
df.insert(13, 'datasetID', nulo)
df.insert(14, 'datasetName', nulo)
df.insert(15, 'ownerInstitutionCode', nulo)
df.insert(16, 'informationWithheid', nulo)
df.insert(17, 'dataGeneralizations', nulo)
df.insert(18, 'dynamicProperties', nulo)

df.to_csv(r'/home/univalle/escritorio/inciva/export
                                    .csv')
```

Después, se validó que todas las columnas del Dataset no tuvieran datos nulos, ya que estos son ruido en cualquier conjunto de datos, así pues, se reemplazaron los elementos nulos con una cadena de texto establecida como "Sin especificar".

3.2. Modelamiento y generación

Posterior al preprocesamiento de la información sobre los especímenes vegetales, se inició el modelamiento de esta en formato RDF. Para esto se creó un Namespace denominado https://sparql.databio.com.co/databio/ para nombrar a la colección biológica y cada uno de los recursos que forman parte de ellas.

En este espacio, los recursos, que se componen de especímenes vegetales se identifican de la siguiente manera https://sparql.databio.com.co/databio/, para identificar a los recursos que forman parte de ella, se debe indicar el nombre de la colección biológica, seguida del símbolo numeral y un fragmento identificador por ejemplo la URI https://sparql.databio.com.co/databio/HerbarioTULV.rdf#TULV%205617 representa al espécimen vegetal Anthurium oxybelium Schott, una especie vegetal del Herbario situado en el Jardín Botánico.

Luego se seleccionaron vocabularios los cuales fueron Darwin Core para describir plantas y SKOS para unir conceptos de diferentes vocabularios. Estos fueron seleccionados porque su estructura dispone de clases y propiedades necesarias para describir los datos. La tabla 1 describe los nombres correspondientes a los vocabularios anteriores.

Tabla 1. NameSpace utilizados

Vocabulario	Prefijo	Namespace
Darwin Core	dwc	http://rs.tdwg.org/dwc/terms/
RDF	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDFS	rdfs	http://www.w3.org/2000/01/rdf-schema#
SKOS	skos	http://www.w3.org/2004/02/skos/core#

Seleccionados los vocabularios, se inició la generación del grafo que contiene la colección biológica representada por medio de un recurso identificado con el https://sparql.databio.com.co/databio/descripcion.rdf y se relaciona con la colección a través de la propiedad skos:related. En este orden de ideas, se produjeron grafos RDF generales para describir los especímenes vegetales, la figura 4 presenta dicho grafo. Es importante mencionar que a partir de un espécimen vegetal, se describieron los demás, puesto que comparten la misma estructura, las cuales se relacionan con el grafo que describe la colección biológica por medio de la propiedad umbel:isAbout.

Ahora bien, para describir un espécimen vegetal, por ejemplo el recurso https://sparql.databio.com.co/databio/HerbarioTULV.rdf#TULV1392 que describe a la especie vegetal *Psychotria grandis Sw*, se usaron propiedades tales como DWC:recordedBy para describir quien o quienes recolectaron dicha especie vegetal, DWC:scientificName para el nombre científico, DWC:family para la familia en la que se clasifica la especie, DWC:institutionCode para el nombre (o acrónimo) usado por la institución que tiene la custodia de los objetos o la información mencionada en el registro, DWC:continent, DWC:country, DWC:stateProvince, DWC:county, DWC:municipality, DWC:locality para el nombre del continente, país, departamento, municipio, corregimiento y predio respectivamente de donde se encontró el espécimen, también se usó DWC:occurrenceID para identificar cada grafo con una especie correspondiente. Las propiedades owl:sameAs y rdfs:seeAlso son utilizadas en todos los grafos para establecer enlaces externos. La figura 5 evidencia todas esas propiedades.

En concreto se generó un Dataset actualmente con 50149 tripletas que describen los especímenes vegetales del Jardín Botánico Juan María Céspedes almacenado en Fuseki y disponible en http://sparql.databio.com.co:3030/databio/. Además, se generó un Endpoint disponible en http://sparql.databio.com.co:3030/databio/sparql para consultas SPARQL.

3.3. Almacenamiento, enlazamiento y publicación

Para almacenar los grafos se serializaron en el formato RDF/XML a través de una librería llamada RDFLib de Python [22], que permite convertir la información codificada a cualquier serialización RDF.

Posteriormente se definió una función para describir a los especímenes vegetales y su modelamiento. Luego, para serializar la información representada por medio de RDFLib se utilizó la función *serialize* y a través del atributo *format* se indica el formato de serialización, que es RDF/XML.

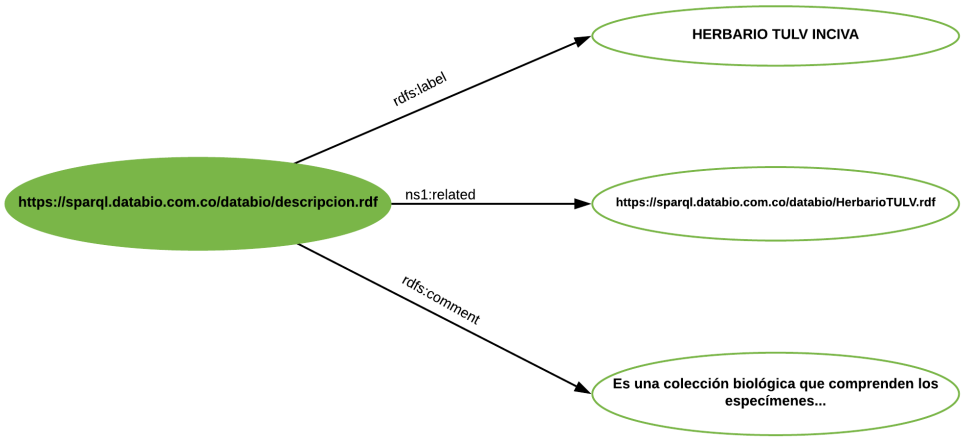

Figura 4. Fragmento del grafo.

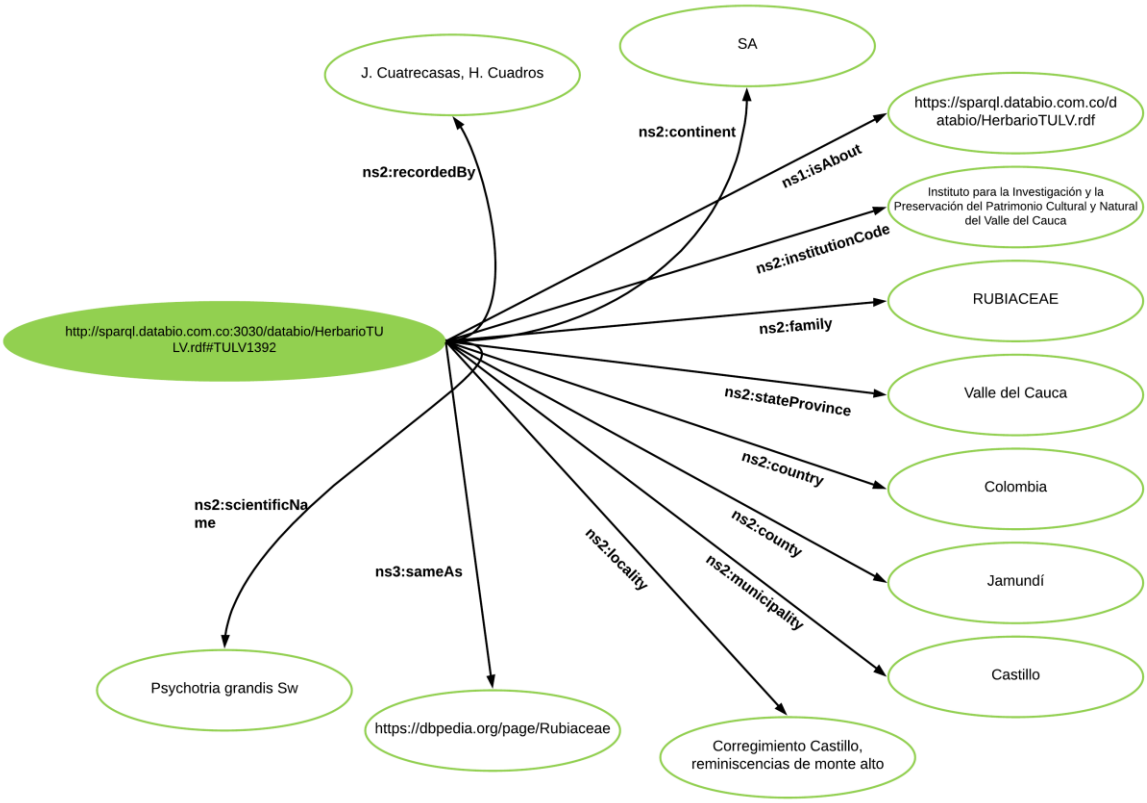

Figura 5. Fragmento del grafo donde se evidencian las propiedades.

Al momento de tener serializada la información ya es posible visualizar los grafos generados, para ello se hizo uso del W3C que posee esta funcionalidad y se encuentra disponible en http://www.w3.org/RDF/Validator/. Esta herramienta resulto útil puesto que permitió validar las tripletas y visualizar a modo de imagen los grafos generados.

Posterior a la serialización, se eligió un triplestore para guardar y consultar la información. En este caso se eligió Apache Jena Fuseki [23], que tiene la capacidad de convertir una consulta SPARQL [24] en una URL para solicitarla a una interfaz REST (Representational state transfer). Para almacenar la información, Fuseki fue instalado en un servidor y se encuentra disponible en la dirección http://sparql.databio.com.co:3030.

Los grafos serializados se enlazaron con recursos presentes en DBpedia y se usó la propiedad owl:sameAs, perteneciente a OWL (Web Ontology Language), la cual permite fusionar las descripciones de los recursos que son equivalentes.

El siguiente fragmento de código representa un espécimen vegetal perteneciente a la colección biológica identificada por medio del URI https://sparql.databio.com.co/databio#TULV5617 que representa a la Anthurium oxybelium Schott, esta es enlazada al recurso http://dbpedia.org/page/Anthurium disponible en DBpedia.

```
<?xml version="1.0" enconding="UTF-8"?>
<rdf:RDF
  xmlns:ns1="http://umbel.org/umbel#"
  xmlns:ns2="http://rs.tdwg.org/dwc/terms"
  xmlns:ns3="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
  syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema"
>
  <rdf:Description
  rdf:about="https://sparql.databio.com.co/
  databio/HerbarioTULV.rdf#TULV5617">
    <ns2:institutionCode>Instituto para
    Investigación y la preservación del Patrimonio
    Cultural y Natural del Valle del
    Cauca</ns2:institutionCode>
    <ns2:county>Tuluá</ns2:county>
    <ns2:country>Colombia</ns2:country>
    <ns2:recordedBy>W. Devia, R. Bernal, F.
    Prado</ns2:recordedBy>
    <ns2:continent>SA</ns2:continent>
    <ns2:municipality>Santa Lucía</ns2:municipality>
```

```
    <ns2:family>ARACEAE</ns2:continent>
    <ns2:scientificName>Anthurium oxybelium Schott
    </ns2:scientificName>
    <ns1:isAbout
    rdf:resource="https://sparql.databio.com.co/
    databio/HerbarioTULV.rdf"
    <ns2:stateProvice>Valle del
    Cauca</ns2:stateProvice>
    <ns2:locality>Corregimiento Santa Lucía, finca
    Las Nieves, v a a La Polonia.</ns2:locality>
    <ns3:sameAs
    rdf:resource="http://dbpedia.org/page/Araceae"
    />
  </rdf:Description>
</rdf:RDF>
```

Finalmente, el Dataset serializado, almacenado y enlazado fue publicado en Datahub [25] un repositorio de datos abiertos enlazados y está disponible en https://datahub.io/ocampo.angelica/databio.

Dentro de los archivos que contiene el Dataset se encuentra el Endpoint, generado por Fuseki, http://sparql.databio.com.co:3030/databio/sparql

En cuanto al Dataset estructurado bajo los principios del LOD, se logró que se actualizará el formato rdf, es decir, los registros de los especímenes vegetales se estructurarán bajo este formato cada que el usuario lo considere necesario, ya que puede descargar la información en esta extensión e implícitamente está generando una actualización del formato que contienen los grafos serializados de los especímenes vegetales.

3.4. Desarrollo de la aplicación

Se diseñó e implementó una aplicación Web bajo la metodología de desarrollo Extreme Programming que tuvo iteraciones durante 34 semanas. La aplicación permite el registro, consulta y descarga de los datos referentes a los especímenes vegetales del jardín botánico (figura 5) en diferentes formatos no propietarios (csv, tsv y rdf). Además de interoperar con la Api de Google Maps y poseer un módulo para usuarios expertos que permite realizar consultas SPARQL al Dataset publicado y enlazado. En términos arquitectónicos se observa la instanciación de una arquitectura de referencia tipo cliente servidor como lo evidencia la figura 6.

Se hizo uso de la base de datos MongoDB para almacenar toda la información, Node.js con Express y Python, para consultar y enviar la información al cliente y Angular 8 para presentar la información consultada en el servidor. Las figuras 7, 8 y 9 son evidencias de la aplicación puesta en producción y actualmente usada por el personal del Jardín Botánico.

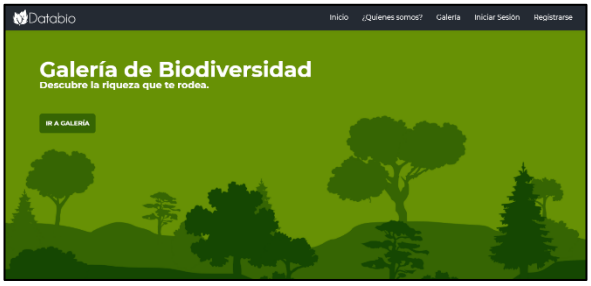

Figura 7. Home de la Aplicación

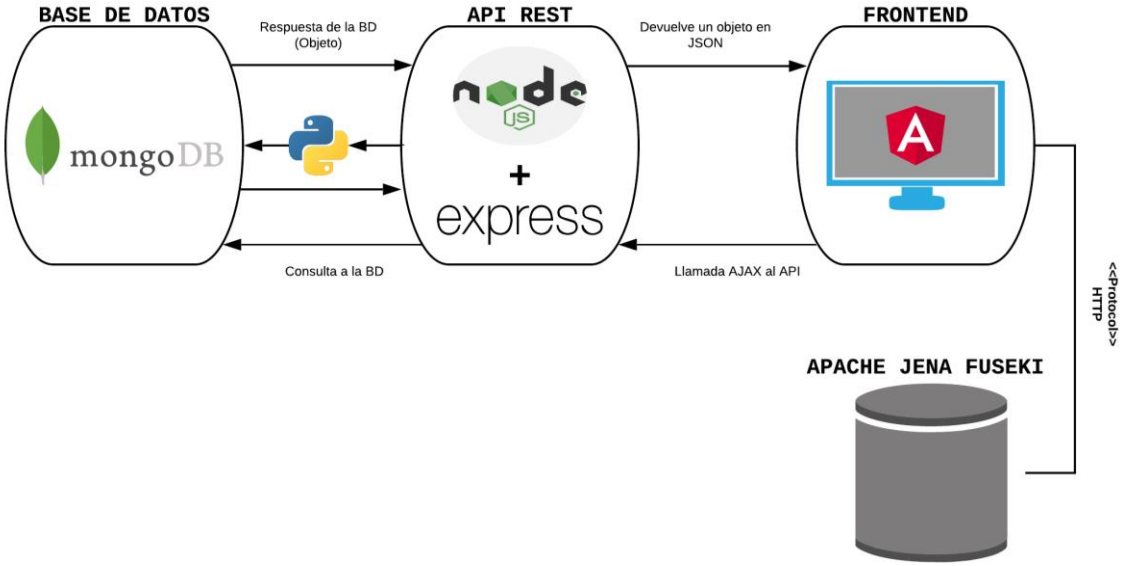

Figura 6. Arquitectura de la Aplicación.

Número de registros: 4182

#	Nombre	Número de Catálogo	Colector(es)	Reino	Phylum	Clase
1	Achatocarpus nigricans Triana	TULV17802	A. Castaño N, W. Devia	Plantae	Magnoliophyta	Magnc
2	Saurauia	TULV17938	A. Castaño N.	Plantae	Magnoliophyta	Magnc
3	Spondias	TULV17964	A. Castaño N.	Plantae	Magnoliophyta	Magnc
4	Unonopsis	TULV7500	W. Devia, R. Bernal, F. Prado	Plantae	Magnoliophyta	Magnc
5	Xylopia	TULV7519	W. Devia, R. Bernal, F. Prado	Plantae	Magnoliophyta	Magnc
6	Xylopia	TULV7520	W. Devia, R. Bernal, F. Prado	Plantae	Magnoliophyta	Magnc
7	Crematosperma	TULV7532	W. Devia, F. Prado	Plantae	Magnoliophyta	Magnc

Figura 8. Visualización de listado de especímenes.

Figura 9. Módulo SPARQL.

La aplicación ofrece información sobre los colaboradores del proyecto disponible en https://databio.com.co/#/aboutUs para facilitar el acuerdo de posibles divulgaciones.

4. Discusión

Teniendo en cuenta que en Colombia hay poca participación en el portal de Datos Abiertos en el contexto de biodiversidad y que no hay iniciativas que incluyan Datos Abiertos Enlazados o Linked Open Data, este trabajo representa un aporte innovador en el contexto de la sistematización de colecciones biológicas, puesto que incluye el modelamiento de información sobre biodiversidad en formato RDF, que permite publicar y enlazar los datos con otros datos presentes en la Web.

Aquí, las universidades y demás actores del ecosistema de ciencia Tecnología e innovación (CTeI), juegan un papel significativo, ya que desde allí se podrían promover proyectos de diferente naturaleza para aportar a la conservación del medio ambiente, por ejemplo, las facultades de ingeniería pueden proporcionar las correspondientes aplicaciones tecnológicas.

De este modo se contribuye a tendencias que destacan que es precisamente la tecnología una oportunidad para la protección y conservación del medio ambiente [26].

En este orden de ideas, se pasó de tener información sobre especímenes vegetales en formato Darwin Core a conseguir las cinco estrellas de datos abiertos esperadas desde el inicio del desarrollo del Dataset, ya que no solo se utilizaron URIs para describir los especímenes vegetales, sino que también se enlazaron con otros datos presentes en la Web Semántica. Es claro que existen muchos repositorios que describen información sobre especies vegetales, dentro de los cuales está GBIF o Uniprot, estos no pudieron incluirse dentro de las URIs que enlazaran la información presente en el Dataset del Jardín Botánico, por poseer un identificador único para cada especie que hizo que el enlazamiento automático desde la aplicación web resultará complejo. Pese a ello, se contó con DBpedia, un repositorio que identifica a las especies por el nombre científico, de esta manera se pudo contemplar la actualización del Dataset desde la aplicación Web.

Databio comparada con herramientas como las presentadas en [7] y [8], se destaca precisamente por facilitar el enlazamiento de los datos con fuentes de terceros, lo cual de manera objetiva apoyara los procesos de toma de decisión en lo que respecta al cuidado de la biodiversidad en el territorio vallecaucano y porque no en otros territorios del país.

Adicionalmente, Databio al ser desarrollada en tecnologías vigentes posibilita una escalabilidad y elasticidad para futuras integraciones con herramientas o incluso la desincorporación de otras, lo cual traduce que es una herramienta que fácilmente puede desempeñarse en contextos diversos lo cual coincide con lo encontrado en [11] que señala que los ecosistemas deben ser capaces de combinar variedad de herramientas para dar soporte al conocimiento abierto en contextos heterogéneos.

Finalmente, un antecedente importante para este trabajo fue Dataeco, un Dataset con información ecoturística del centro del Valle del Cauca. Este fue implementado bajo los principios del LOD. Los resultados de Databio en contraste con Dataeco evidencia que el primero contempló la actualización del formato rdf con todos los especímenes vegetales que se encuentren registrados en la aplicación, incluyendo el enlazamiento de cada uno con su respectivo recurso presente en DBpedia. En este orden de ideas, un cambio en la base de datos genera nuevamente el grafo RDF, siendo este, un método para la actualización constante del Dataset.

5. Conclusiones

El producto aquí presentado contribuye en algún grado a políticas de gobierno como por ejemplo la Ley 1712 de 2014 sobre Transparencia y del Derecho de Acceso a la Información Pública Nacional. También con políticas o acuerdos internacionales, un ejemplo de ello son los Objetivos de Desarrollo Sostenible (ODS) de la Organización de las Naciones Unidas (ONU), específicamente los objetivos 13 y 15, referente a Acción por el Clima y Vida de Ecosistemas Terrestres respectivamente.

Databio es la base para seguir reforzando el proceso de sistematización de colecciones vegetales, es decir aumentar el valor entregado a los diferentes grupos de interés, siempre y cuando exista un proceso continuo de mantenimiento y desarrollo para dotar al sistema de capacidades funcionales como sistemas de alertas tempranas para especies en peligro, modelos estadísticos, sistemas de recomendación para comunidades expertas, minería de datos entre otros.

Aunque las instituciones responsables de la conservación de la biodiversidad en el país, publican en la web información estandarizada en formato Darwin Core, por medio de los IPT, el hecho de enlazar la información con repositorios de terceros se traduce en un componente innovador en ese contexto, e implica asumir aún más responsabilidades para promover e incentivar desarrollos tecnológicos como Databio, que seguramente una vez estabilizado en ambiente productivo apoyara no solo la sistematización de diferentes tipos de colecciones sino que también apoyara los procesos de divulgación.

El diseño arquitectónico de Databio facilita la interoperabilidad con otros sistemas, por ejemplo, con el Sistema Nacional Ambiental (SINA) que es administrado por la red nacional de datos abiertos sobre biodiversidad. Sin embargo, también depende del continuo mantenimiento y la voluntad de trabajar conjuntamente con enfoques de asociatividad. Este tipo de iniciativas implica pensar en estrategias como Data as a Service con claros enfoques orientados a la calidad de los datos.

Agradecimientos

Databio en su primera versión fue posible gracias a Alejandro Castaño Naranjo, curador de colecciones vegetales de INCIVA, su acompañamiento fue fundamental para avanzar en dirección correcta referente a la sistematización, enlazamiento y publicación de especímenes vegetales del Jardín Botánico del Valle Juan María Céspedes. Por supuesto, nuestro agradecimiento a INCIVA, quien, además de su participación técnica aportó los recursos para la puesta en producción por un año de la aplicación.

Referencias

[1] R. Luna, A. Castañon, A. Raz-Guzmán, “La biodiversidad en México: su conservación y las colecciones biológicas”, *Ciencias UAM*, no. 101, pp. 8, 2011.

[2] D. Vélez et al., “Difusión de datos biológicos en la red como apoyo a la educación ambiental, investigación científica y conservación de la biodiversidad en Colombia”, *Renata*, vol. 2, no. 4, pp. 49-57, 2012.

[3] K. Giancarlo, B. Acosta, “Sistematización Colecciones biológicas IAvH”, *Repositorio Institucional de Documentación Científica*, 2017 [En línea]. Disponible en: http://hdl.handle.net/20.500.11761/34074

[4] GBIF, “IPT: Herramientas para la publicación integrada de datos”, 2015 [En línea]. Disponible en: https://www.gbif.org/es/ipt

[5] SIB, "Colecciones en Línea", 2020 [En línea]. Disponible en: https://colecciones.biodiversidad.co/

[6] C. Castrillón -Arias, C. A. Agudelo-Henao, O. A. Vega, "Plataforma Web para Colecciones Biológicas: Caso Herbario Universidad del Quindío", *Sci. Tech.*, vol. 23, no. 2, pp. 249-257, 2018.

[7] Y. Alvarado, J. Fernández, R. Guerrero, G. Rodríguez, N. Jofré, "Plataforma para Repositorios Digitales 3D de Colecciones Biológicas", en *XXV congreso argentino de ciencias de la computación*, 2019, pp. 428-437.

[8] S. Ortega, A. Guevara, "Darwin Core: Estándar para la Gestión de Datos Biológicos Primarios en la UTN", en *Encuentro latinoamericano Eciencia 2017,* pp. 1-19.

[9] INCIVA, "Quiénes somos La Institución INCIVA - Instituto para la Investigación y la Preservación del Patrimonio Cultural y Natural del Valle del Cauca", 2020 [En línea]. Disponible en: http://www.inciva.gov.co/institucion/quienes-somos

[10] M. León Santos, G. Ponjuán Dante, M. Rodríguez Calvo, "Procesos estratégicos de la gestión del conocimiento", *Acimed,* vol. 14, no. 2, 2006.

[11] A. García Holgado, F. J. García - Peñalvo, "Gestión del conocimiento abierto mediante ecosistemas tecnológicos basados en soluciones "open source"", en *Ecosistemas del Acceso Abierto,* Salamanca, España: Ediciones Universidad de Salamanca, 2018, pp. 147-160.

[12] C. Gonzalo, "Ontologías y la Web Semántica", 2005 [En línea]. Disponible en: https://www.carlosgonzalo.es/ontologias-y-la-web-semantica/

[13] Z. Ma, M. A. M. Capretz, L. Yan, "Storing massive Resource Description Framework (RDF) data: A survey", *Knowl. Eng. Rev.*, vol. 31, no. 4, pp. 391-413, 2016, doi: 10.1017/S0269888916000217

[14] M. Lnenicka, J. Komarkova, "Big and open linked data analytics ecosystem: Theoretical background and essential elements", *Gov. Inf. Q.*, vol. 36, no. 1, pp. 129-144, 2019, doi: 10.1016/j.giq.2018.11.004

[15] H. Firmino, G. Lima, "Reuso de ontologias: uma revisão de literatura no contexto da web semântica", en *Tendências atuais e perspetivas futuras em organização do conhecimento*, 2017, pp. 487-497.

[16] TDWG, *DarwinCore: una guía de referencia rápida, versión 3.0.* Bogotá: SiB Colombia, 2018.

[17] G. de Colombia, "Datos Abiertos - Gobierno Digital", *Ministerio de Tecnologías de la Información y las Comunicaciones*, 2014 [En línea]. Disponible en: https://estrategia.gobiernoenlinea.gov.co/623/w3-article-9407.html.

[18] INCIVA, "Ubicación Jardín Botanico INCIVA - Instituto para la Investigación y la Preservación del Patrimonio Cultural y Natural del Valle del Cauca", 2020 [En línea]. Disponible en: http://inciva.gov.co/patrimonio-turistico/jard-iacute-n-bot-aacute-nico-juan-maria-c-eacute-spedes

[19] J. Chavarriaga, H. Arboleda, G. Lidis, "Modelo de Investigación en Ingeniería del Software: Una propuesta de investigación tecnológica", *Ing. del Softw. y Sist.*, 2004.

[20] V. Zuluaga, C. Moreno, "Creación de un dataset sobre ecoturismo de los municipios de Riofrío y Tuluá para publicar en la Web de Datos", tesis de grado, Universidad del Valle, 2016.

[21] R. Moya, "Pandas en Python, con ejemplos -Parte I- Introducción", Jarroba, 2015 [En línea]. Disponible en: https://jarroba.com/pandas-python-ejemplos-parte-i-introduccion/

[22] RDFLib Team, "rdflib 5.0.0 - rdflib 5.0.0 documentation", *RDFLib*, 2002 [En línea]. Disponible en: https://rdflib.readthedocs.io/en/stable/

[23] Apache Jena, "Apache Jena - Apache Jena Fuseki", 2011 [En línea]. Disponible en: https://jena.apache.org/documentation/fuseki2/

[24] SKOS Simple Knowledge Organization System Primer, "SPARQL Lenguaje de consulta para RDF", *W3C*, 2009 [En línea]. Disponible en: https://skos.um.es/TR/rdf-sparql-query

[25] Power Data, "¿Qué es Datahub, Data Lake y Datawarehouse?", 2018 [En línea]. Disponible en: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-es-datahub-data-lake-y-datawarehouse

[26] Portafolio, "La tecnología, un 'salvavidas' para preservar la biodiversidad | Opinión | Portafolio", *Otros columnistas,* 2020 [En línea]. Disponible en: https://www.portafolio.co/opinion/otros-columnistas-1/la-tecnologia-un-salvavidas-para-preservar-la-biodiversidad-538683