UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

METADATOS, ESTÁNDARES Y DATOS LIGADOS EN COLECCIONES DIGITALES DE DIVERSIDAD BIOLÓGICA

T E S I S

QUE PARA OBTENER EL TÍTULO DE :

BIÓLOGA

P R E S E N T A:

NORMA GUADALUPE RODRÍGUEZ LUIS

DIRECTORA DE TESIS

DRA. LAYLA MICHÁN AGUIRRE

CIUDAD UNIVERSITARIA, CD. MX., 2019

HOJA DE DATOS DEL JURADO

1. Datos del alumno
Rodríguez
Luis
Norma Guadalupe
5515570859
Universidad Nacional Autónoma
de México
Facultad de Ciencias
Biología

2. Datos del tutor
Dra.
Layla
Michán
Aguirre

3. Datos del sinodal 1
Dra.
Patricia
Koleff
Osorio

4. Datos del sinodal 2
Dr.
Erick Alejandro
García
Trejo

5. Datos del sinodal 3
Dr.
César
Ríos
Muñoz

6. Datos del sinodal 4
Dra.
Graciela
Zamudio
Varela

7. Datos del trabajo escrito
Metadatos, estándares y datos ligados en colecciones digitales de diversidad biológica
94 pp.
2019

Agradecimientos

Agradezco a la Universidad Nacional Autónoma de México por haberme permitido formarme como profesionista. A mi tutora, la Dra. Layla Michán Aguirre por ser mi guía en esta etapa, por su ayuda, paciencia, apoyo, consejos y sobre todo por compartir conmigo su conocimiento y visión interdisciplinaria de la biología. Así mismo, quiero expresar mi sincero agradecimiento a mis sinodales Dra. Patricia Koleff Osorio, el Dr. Erick Alejandro García Trejo, Dr. César Ríos Muñoz y Dra. Graciela Zamudio Varela, por su tiempo, atenciones, comentarios y aportaciones que mejoraron significativamente esta tesis.

Agradezco a los miembros del laboratorio Lidia, Diana y Roberto por sus consejos y comentarios para mejorar este trabajo. A todos los profesores que me alentaron y enseñaron a aprender y superarme cada día; a mis profesores de la Facultad de Ciencias por contagiarme su pasión y amor por estudiar y comprender la vida y al Fis. Martín Alarcón Ronzón por ayudarme a aprender matemáticas para biólogos.

Por otro lado, y sin restar importancia, agradezco a todas las personas involucradas en esta etapa de mi vida en mi formación como ser humano. Principalmente a mi familia por apoyarme siempre, a mis padres José y Margarita por cuidarme, darme la oportunidad de estudiar, brindarme todo su apoyo e inculcarme valores como la perseverancia. A mi hermano Misael, tíos y abuelos por toda su ayuda incondicional.

Agradezco también a Paola V. Jiménez por su gran amistad, consejos, paciencia y ayuda en momentos difíciles. A Carlos Orozco Alarcón, por su compañía, por motivarme a culminar este ciclo y por sus observaciones y cariño. A Anahí Camarena por su valiosa amistad, risas y platicas. A mis amigas Cecilia Sandoval, Janeth Núñez, Surya González y Lizbeth Santiago. A Tania, Griselda, Jorge, Omar, Agustín, Angélica y Lidia. A Martha Orozco y a todas aquellas personas con las que compartí momentos especiales y de las que aprendí algo en este proceso, dentro y fuera de la escuela. Gracias.

Dedicada a la memoria de José Carlos Rodríguez Luis,

gracias por enseñarme tanto de la vida.

A mi familia.

Índice de contenido

Índice de figuras

Índice de tablas

Resumen

INTRODUCCIÓN: Las bases de datos disponibles en la *World Wide Web* para la publicación y almacenamiento de datos en biodiversidad presentan problemas de compatibilidad que representan barreras para la búsqueda, reutilización, exploración, análisis e interpretación de la información a distintas escalas.

OBJETIVOS: Describir el estado del conocimiento respecto al desarrollo e implementación de metadatos, estándares y datos ligados en el manejo de datos mundiales de especies en colecciones digitales, a partir de los artículos de investigación y las bases de datos más relevantes.

MÉTODOS: Se realizó la búsqueda, recuperación y selección de la literatura científica más relevante sobre el manejo de datos de especies a escala mundial. Los documentos seleccionados (95), se analizaron con las siguientes variables: colecciones digitales de diversidad biológica a nivel de especies, esquemas de metadatos, estándares, identificadores, sistemas de organización del conocimiento y tecnologías afines a los datos ligados. Por último, se examinó el uso de metadatos, estándares y datos ligados en las colecciones digitales que indexan datos de todos los grupos de seres vivos. El método incluyó la creación de una base de datos para integrar y sistematizar la información del proyecto.

RESULTADOS: Se identificaron 72 colecciones digitales de especies, el mayor número descrito formalmente hasta el momento; y 131 recursos web para el manejo de información digital, de los cuales 46 son específicos para el manejo de datos de organismos. Las colecciones digitales más destacadas fueron la *Global Biodiversity Information Facility* y *Catalogue of Life* respecto al número de citas y enlaces a otras bases de datos. Por otra parte, los recursos digitales creados en biodiversidad son principalmente sistemas de organización del conocimiento de tipo ontología y estándares para transferir información entre colecciones. La especificación *Resource Description Framework* fue el recurso digital más relevante relacionado con la publicación de datos ligados de especies. Finalmente, se obtuvo que la mayoría de las colecciones digitales de diversidad biológica implementan esquemas de metadatos, estándares e identificadores comunes, con deficiencias en el uso de sistemas de organización del conocimiento compartidos.

CONCLUSIÓN: El estado del conocimiento respecto al uso y desarrollo de metadatos, estándares y datos ligados en el manejo de datos de diversidad biológica a nivel de especies, evidencia la falta de bases de datos y recursos digitales para gestionar algunos dominios y tipos de información, la ausencia de consensos en la adopción de las herramientas existentes y el creciente interés en la publicación de datos abiertos y datos ligados en la web semántica.

PALABRAS CLAVE: Datos; Informática de la biodiversidad; web 3.0; acceso abierto; identificadores; sistemas de organización del conocimiento (SOC)

Introducción

La biología moderna se encuentra en una era de acumulación acelerada de información, donde los científicos dependen de grandes cantidades de datos almacenados en medios digitales. Este crecimiento exponencial de datos biológicos, producto de la revolución informática, ha traído consigo nuevas formas de generar, sistematizar, analizar y transmitir la información (Howe *et al*., 2008; Michán, 2011).

Uno de los principales desafíos para el manejo de información digital, es la curación, gestión o manejo de datos, que consiste en la administración y conservación de los datos, con el fin de que permanezcan confiables y disponibles para su reutilización por cualquier usuario (Pennock, 2007). En esta tesis de tipo documental, se describe el estado del conocimiento respecto al desarrollo y uso de recursos digitales en la gestión de datos mundiales de especies en bases de datos de la *World Wide Web*.

Las diversas disciplinas que buscan comprender la extensión y trayectoria de la vida en la Tierra, han acumulado una gran cantidad de datos que cumplen con las características de los datos masivos o *big data*, por su volumen, variedad y velocidad de acumulación (Bowker, 2000; Schnase *et al*., 2003; Beckstein *et al*., 2014). Tan solo en la base de datos *Global Biodiversity Information Facility* se han reportado hasta el año 2018 mil millones de registros que representan la presencia de especies (GBIF.org, 2018a). La magnitud y aspectos de la diversidad biológica representan desafíos para la integración y uso de los datos de especies por lo que se requieren de investigaciones centradas en datos dentro de las ciencias de la información (Edwards, Lane y Nielsen, 2000; Wieczorek *et al*., 2012).

Las bases de datos o colecciones digitales son las herramientas fundamentales para resguardar los datos de biodiversidad, ya que en ellas se almacena, organiza, integra y distribuye la información de especies; además de ser instrumentos de aprendizaje y un medio para la colaboración que acelera el proceso de las investigaciones científicas (Turnhout y Boonman-Berson, 2011; SeaLifeBase, 2018). Una colección digital de diversidad biológica, se define en esta tesis como un conjunto estructurado de datos y metadatos de especies u organismos, estandarizado, curado e indexado de forma taxonómica, que se encuentra disponible en línea para su consulta interactiva.

Actualmente diversas iniciativas tienen como objetivo almacenar y publicar datos mundiales de diversidad biológica, lo que ha provocado dificultades para la integración, intercambio y comparación de la información disponible en línea (Hoffmann *et al.*, 2011). Los problemas de compatibilidad entre datos de especies que provienen de distintas fuentes, son

persistentes a pesar del desarrollo de recursos web que facilitan la unificación de criterios para la indexación de la información, como repositorios, ontologías, estándares, vocabularios y servicios web (Koureas *et al*., 2016; Parr y Thessen, 2018). La interoperabilidad entre las colecciones digitales de biodiversidad puede comenzar a habilitarse al conocer el estado actual de los estándares y protocolos disponibles para representar la información (Jonhson, 2007; Daltio y Medeiros, 2008; Goddard *et al*., 2011).

Esta tesis se realizó dentro de la informática de la biodiversidad, una disciplina con el objetivo de emplear técnicas informáticas en la gestión, presentación, descubrimiento, exploración algorítmica, análisis e interpretación de datos de biodiversidad, principalmente a nivel de especies (Soberón y Peterson, 2004; Johnson, 2007; Martellos y Attorre, 2012). Como se observa en la figura 1, este trabajo se realizó específicamente dentro de la interdisciplina entre la biología, la informática y la gestión de datos.

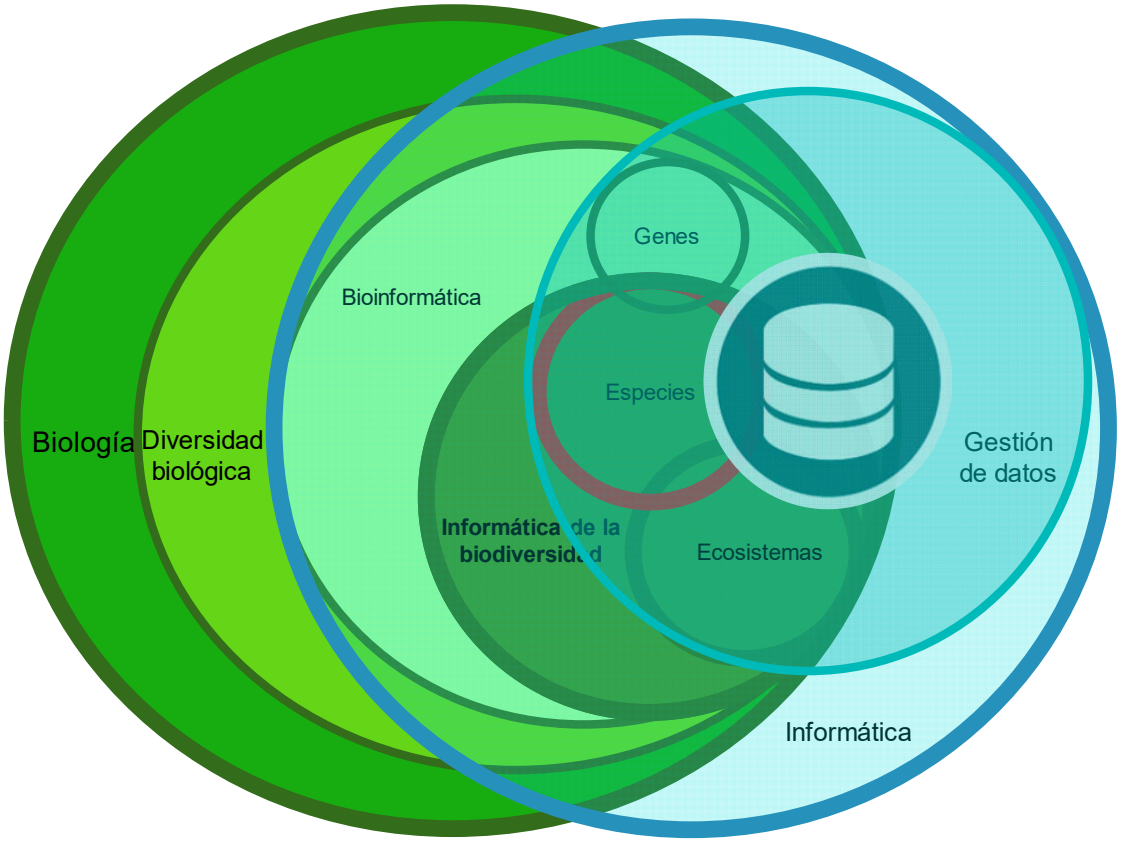


Figura 1. Dominios, áreas, temas y objetos de estudio (Elaboración propia).

Antecedentes

En esta sección se presentan los conceptos básicos sobre el manejo de datos digitales de especies, sus relaciones (figura 2) y las investigaciones previas más sobresalientes.

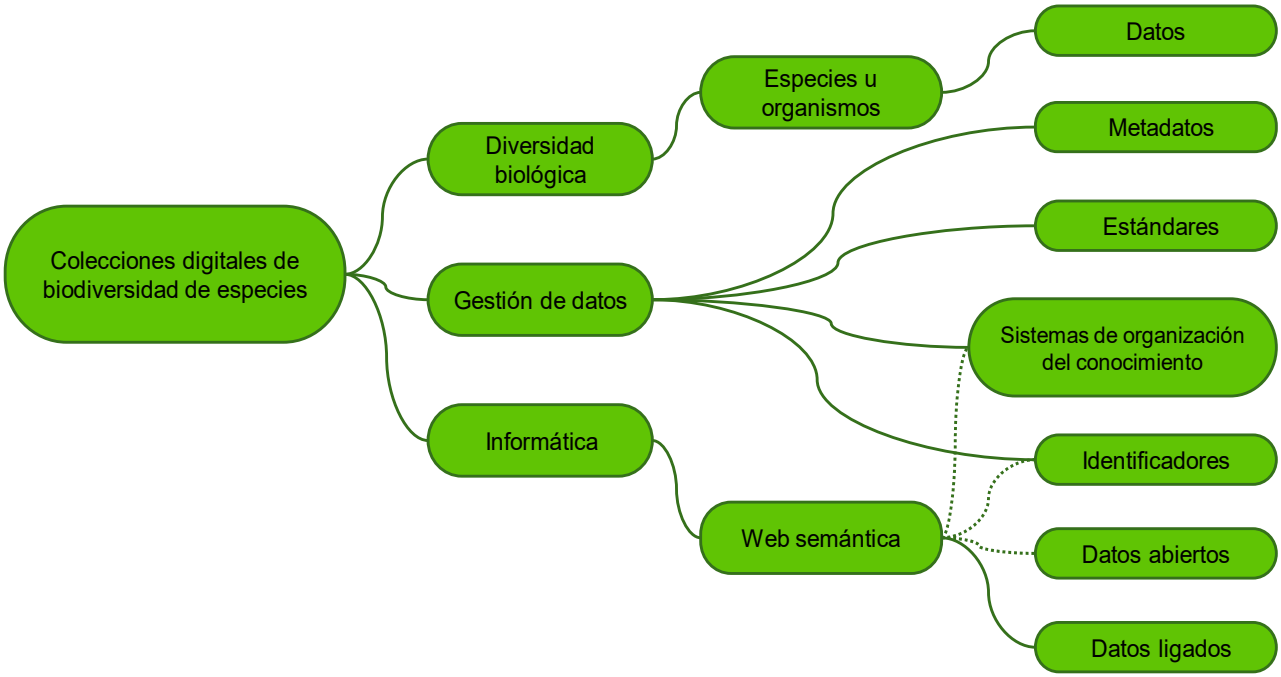


Figura 2. Marco conceptual del estudio (Elaboración propia).

Colecciones digitales

Las colecciones digitales, bases de datos o repositorios, son conjuntos de datos estructurados y relacionados entre sí, que se almacenan y mantienen dentro de sistemas informáticos. Los sistemas informáticos, a su vez, se conforman por los datos, el gestor de base de datos, las aplicaciones web, el equipo informático y los usuarios (Johnson, 2007; Foster y Godbole, 2016). De acuerdo con Foster y Godbole (2016) las principales ventajas de las bases de datos son:

1. Evitar inconsistencias de información
2. Compartir los datos almacenados
3. Eliminar redundancias
4. Facilitar la visualización de los datos
5. Optimizar el mantenimiento y recuperación de información

6. Reducir la velocidad de procesamiento de los datos
7. Resguardar la seguridad de la información
8. Estandarizar la información

Los repositorios que integran datos mundiales de especies, han sido también denominados bases de datos sinópticas (Bowker, 2000), recursos de datos de biodiversidad (Canhos *et al*., 2004), sistemas de información sobre biodiversidad (Daltio y Medeiros, 2008), agregadores de bases de datos (Santos y Branco, 2012), plataformas de megaciencia (Triebel, Hagedorn y Rambold, 2012), centros de datos (Costello y Wieczorek, 2014) y redes de datos a gran escala (Franz y Sterner, 2018). Estas colecciones, integran y preservan a largo plazo información de organismos que proviene de múltiples sitios y disciplinas (Bowker, 2000).

La infraestructura que permite el descubrimiento y análisis de información biológica a gran escala, se conforma por nodos regionales o nacionales, que a su vez integran datos de instituciones, proyectos e individuos (Holetschek *et al*., 2009; Xiao-Ting, 2012; De Pooter *et al*., 2017). En nuestro país, el Sistema Nacional de Información sobre Biodiversidad (SNIB) de México, es el encargado de la compilación, organización y distribución de la información biológica; una actividad estratégica para el manejo y conservación de los recursos naturales en un país megadiverso (CONABIO, 2008). Los datos procedentes de colecciones biológicas científicas y bancos de información de diversas instituciones de investigación y enseñanza, se reúnen en la Red Mundial de Información sobre Biodiversidad (REMIB) que puede ser consultada libremente por cualquier persona desde cualquier lugar del mundo (CONABIO, 2015).

Edwards, Lane y Nielsen (2000) describen la creación de iniciativas en la *World Wide Web* para la distribución de datos de especies como la *Global Biodiversity Information Facility* (GBIF) y *Species2000*. Turnhout y Boonman-Berson (2011) destacan el papel de las bases de datos en la globalización de la biodiversidad y reconocen la falta de integración entre las actividades de investigación y la recopilación de datos a gran escala. Triebel, Hagedorn y Rambold (2012) identifican problemas en la curación de datos y en el desarrollo de programas informáticos a largo plazo, dentro de las colecciones digitales de diversidad biológica. Finalmente, Bingham *et al*. (2017) evalúan las conexiones entre repositorios, servicios web e iniciativas a escala mundial y regional, lo que representa uno de los estudios más completos y relevantes que busca dilucidar el panorama actual de la informática de la biodiversidad.

Diversidad biológica

La diversidad biológica o biodiversidad, es la variedad de seres vivos presentes en el planeta Tierra. Dentro de la biodiversidad se distinguen tres principales escalas: la diversidad genética, la diversidad de organismos o especies; y la diversidad de ecosistemas (CDB, 1992). Tras el Convenio sobre la Diversidad Biológica en el año 1992, la biodiversidad se ha convertido en un tema central de las investigaciones científicas, que buscan caracterizar y determinar sus tendencias a futuro a través de extrapolaciones y proyecciones (Turnhout y Boonman-Berson, 2011).

El conocimiento de la diversidad biológica a nivel de organismos o especies, puede dividirse en tres principales dominios de información de acuerdo con Triebel, Hagedorn y Rambold (2012): taxonomía-nomenclatura, donde se incluyen datos de nombres y clasificación; representatividad geográfico-temporal, que indica la presencia de un organismo en un lugar y tiempo determinado (Holetschek *et al*., 2009); y rasgos-datos descriptivos, que representan las cualidades de los organismos (GEOBON, 2019).

La biodiversidad, es uno de los dominios de conocimiento más importantes para la humanidad, con un amplio volumen de datos a diferentes escalas temporales, espaciales y disciplinares, que pueden ser utilizados con un sinfín de propósitos principalmente científicos, gubernamentales, educativos y comerciales (Bowker, 2000; Schnase *et al*., 2003; Peterson *et al*., 2010; Turnhout y Boonman-Berson, 2011). Autores como Bluhm, Watts y Huettmann (2010) y Turnhout y Boonman-Berson (2011) reconocen la necesidad de bases de datos de acceso abierto, como un medio para almacenar y distribuir información digital de especies en la *World Wide Web*.

Gestión de datos

La gestión de datos, también denominada manejo de datos o curación, es el desarrollo de arquitecturas, políticas, prácticas y procedimientos para el manejo de la información. En estas actividades se incluye la administración de datos en archivos como bancos de datos, bases de datos, portales web y otros recursos digitales similares (Ison *et al.*, 2013). La gestión de datos, se encuentra directamente relacionada con la interoperabilidad, es decir, con la capacidad de intercomunicación, intercambio e interpretación de información entre dos o más sistemas, usuarios u organizaciones independientes con intereses comunes; una condición necesaria para las actividades de investigación y las políticas que agrupan datos de biodiversidad (Groom *et al*., 2017; Ahmad *et al*., 2019).

Uno de los tipos de herramientas para normalizar datos digitales, son los recursos web que forman parte de la Internet y pueden ser localizados mediante cualquier versión del protocolo HTTP (*Hypertext Transfer Protocol*), (Berners-Lee *et al*., 1998; W3C, 1999). Dentro de estos instrumentos destacan los esquemas de metadatos, estándares, sistemas de organización del conocimiento (SOC) e identificadores.

La curación en biodiversidad tiene el objetivo de organizar, estandarizar, normalizar, clasificar y anotar la información para su análisis (Castillo, Michán y Martínez, 2014). Dentro de los estudios en el tema, Goddard *et al*. (2011) analizan la gestión de datos bajo un enfoque sociológico y tecnológico; entre sus propuestas se encuentran la creación de una comunidad de personas especializadas en el alojamiento y archivo de datos, establecer instancias para alojar los datos y desarrollar herramientas para el descubrimiento de la información. Costello y Wieczorek (2014) enlistan las prácticas para la gestión de datos de biodiversidad, donde destacan el uso de estándares, la publicación de datos dentro de colecciones digitales, el acceso abierto y el uso de metadatos para la posterior interpretación de la información. Por su parte, Hugo y colaboradores (2016) recalcan el papel de la publicación de información para resolver los problemas de descubrimiento y acceso en los datos de especies.

Metadatos

Los metadatos son datos asociados a un objeto o recurso, que describen sus características de contenido, ubicación, calidad, creación, etc. (Woolcott, 2017; ANDS, 2018). Proporcionan información sobre el alcance la procedencia, los métodos, las limitaciones, las unidades, los formatos, la licencia de uso y la información de contacto; entre sus funciones principales se encuentran: describir, explicar, localizar y facilitar la reutilización de los datos (Groom *et al*., 2017; Woolcott, 2017). Dentro de los metadatos, los recursos digitales que especifican la descripción y estructuración de un tipo particular de datos se conocen como esquemas de metadatos, estándares de metadatos o bien metadatos descriptivos estructurados (Thessen y Patterson, 2011; Willis, Greenberg y White, 2012).

En el manejo de información científica, los metadatos normalizan la documentación de datos y son por lo tanto, un elemento clave en la publicación, búsqueda, citación y reutilización de la información presente en bases de datos (Willis, Greenberg y White, 2012; ANDS, 2018). Dentro de las ciencias biológicas los metadatos permiten optimizar la anotación e interpretación de la información, y en biodiversidad se reconoce su papel en la integración de diversos conjuntos de datos de especies (Costello *et al.*, 2013; Vos *et al*., 2014; Reimer *et al*., 2017). En

este sentido, el intercambio de datos en biodiversidad requiere de un consenso para determinar los metadatos que deben ser capturados y conservados a largo plazo (Brainerd *et al*., 2017).

Entre las principales investigaciones en el tema, destacan las aportaciones de Wiezoreck *et al*. (2012) quienes presentan la evolución y desarrollo del esquema de metadatos *Darwin Core*, diseñado para compartir datos de especies a partir de la *Dublin Core Metadata Initiative* (DCMI). Así mismo, Walls *et al*. (2014) caracterizan el panorama actual de los esquemas de metadatos y ontologías en biodiversidad, desde una perspectiva semántica. Senderov, Georgiev y Penev (2016) proponen un modelo para optimizar el flujo de trabajo en conjuntos de metadatos. Méndez-Muñoz *et al*. (2017) describen una guía para la gestión de metadatos en colecciones de biodiversidad. Finalmente, Tessarolo *et al*. (2017) destacan la degradación y pérdida de datos y metadatos debido a su manejo inadecuado.

Estándares

Los estándares, también denominados normas o especificaciones, son pautas establecidas por consenso u organizaciones reconocidas, para codificar e intercambiar datos comunes (NCIT, 2019). Describen el origen, la calidad y uso potencial de la información y facilitan la gestión de los datos, al reducir su pérdida y duplicación (Hoffmann *et al*., 2011; Vos *et al*., 2014; Brainerd *et al*., 2017). El uso de estándares y protocolos en biodiversidad, proporciona interoperabilidad e integración a los datos provenientes de distintos sistemas distribuidos en la *World Wide Web* (Costello y Vanden Berghe, 2006).

El papel de los estándares en el manejo de datos de especies, es descrito por Halpin *et al*. (2006) quienes destacan su uso para el correcto diseño y desarrollo de las colecciones digitales. Johnson (2007) realiza una de las revisiones más completas acerca del desarrollo e implementación de estándares en la gestión, integración, intercambio y análisis de datos de presencia de especies y caracteres, desde una perspectiva entomológica. Costello *et al*. (2013) examinan la adopción de estándares para controlar la calidad y validar la publicación de datos de biodiversidad. Zermoglio, Guralnick y Wieczorek (2016) proponen el desarrollo de estándares para la validación de nombres de taxones. Finalmente, Groom *et al*. (2017) identifican la falta de estandarización en la información en biodiversidad y la falta de estándares que describen los atributos de las especies exóticas. Dentro de la informática de la biodiversidad la institución encargada del desarrollo, ratificación y promoción de los estándares para el intercambio de datos a nivel de organismos es la *Biodiversity Information Standards* (TDWG, 2018).

Sistemas de organización del conocimiento (SOC)

Los sistemas de organización del conocimiento, son un tipo de especificación que modela la estructura semántica implícita en un dominio de información, con el uso de etiquetas, definiciones, relaciones y propiedades de conceptos (Zeng, 2008; Woolcott, 2017; Mazzocchi, 2018). Los SOC, se utilizan para anotar y estructurar datos, facilitan la búsqueda, interpretación, intercambio y recuperación de contenido digital por humanos o máquinas; además de que permiten visualizar el alcance, las relaciones semánticas y las necesidades de información de una base de datos (Hodge, 2000; Catapano *et al*., 2011; Koureas *et al*., 2016; Rosati *et al*., 2017).

Los elementos básicos que componen a los sistemas de organización del conocimiento son: los términos o conceptos, las definiciones y en los casos más sofisticados las conexiones semánticas entre los términos (Bratková y Kučerová, 2014). La clasificación y tipos de SOC se detallan a continuación (tabla 1).

Tabla 1

Clasificación y tipos de sistemas de organización del conocimiento (SOC) de acuerdo con Hodge (2006)

Categoría	Descripción	Tipo SOC	Ejemplos
Listas de términos	Conjuntos de términos que incluyen definiciones (en la mayoría de casos)	Archivos de autoridad	*Library of Congress Name Authority File*
		Diccionarios	----------
		Diccionarios geográficos	*U.S. Code of Geographic Names*
		Glosarios	*Environmental Protection Agency (EPA)*
Clasificaciones y categorías	Conjuntos de términos enfocados en temas o materias	Esquemas de categorización	*Medical Subject Headings (MeSH)*
		Esquemas de clasificación, taxonomías y esquemas de categorización	*Dewey Decimal Classification*

Tabla 1 (continuación)

Categoría	Descripción	Tipo SOC	Ejemplos
Listas de relaciones	Conjuntos de términos que representan conceptos y sus relaciones	Tesauros	*Unified Medical Language System (UMLS)*
		Redes semánticas	*Princeton University's WordNet*
		Ontologías	*Gen Ontology*

Fuente: Modificado de Mazzocchi (2018).

En biodiversidad Walls *et al*. (2014) describen de forma general el desarrollo y adopción de ontologías para la descripción de organismos, muestras ambientales y observaciones. Recientemente, Parr y Thessen (2018) discuten el uso de ontologías y vocabularios desde la perspectiva del usuario y destacan la necesidad de implementar tecnologías semánticas dentro de las colecciones digitales.

Identificadores

Los identificadores, son caracteres generalmente numéricos, asignados para reconocer y localizar objetos almacenados en archivos, bases de datos relacionales, aplicaciones u otras fuentes de información digital (Ison *et al*., 2013; LSID, s.f.). Tienen el potencial de distinguir de forma inequívoca a los registros de especies y proporcionan un enlace entre información relacionada; facilitan el descubrimiento de los datos y refuerzan los enlaces entre ellos (Nelson, Sweeney y Gilbert, 2018). Idealmente, los identificadores presentes en la *World Wide Web* deben ser únicos y persistentes, ya que representan un vínculo permanente a los datos y metadatos, a pesar de los cambios en su ubicación (ANDS, 2018).

En biodiversidad, Page (2006) examina detalladamente el uso de identificadores persistentes y únicos a nivel mundial como el *Digital Object Identifier* (DOI), *Uniform Resource Identifier* (URI), *Hypertext Transfer Protocol* (HTTP) y *Life Sciences Identifier* (LSID) para vincular datos de interés en las investigaciones. Guralnick y colaboradores (2015) destacan la importancia de los identificadores únicos y globales asociados a datos y muestras en biocolecciones para la recopilación y acceso a la información. Güntsch *et al*. (2017) proponen el uso de identificadores compatibles con los datos abiertos vinculados, para satisfacer las actividades de investigación

basadas en la web semántica. Finalmente, Nelson, Sweeney y Gilbert (2018) recomiendan el uso de identificadores a nivel global para vincular especímenes de herbarios y destacan su importancia en la gestión de datos disponibles electrónicamente.

Datos ligados

El componente central de la Internet es la *World Wide Web*, una red creada para publicar y compartir información a partir de tres tecnologías básicas: *HyperText Markup Language* (HTML), *Uniform Resource Identifier* (URI) e *Hypertext Transfer Protocol* (HTTP). La web 3.0 o web semántica, es una iniciativa que busca crear una red de datos estructurados, con base en la conexión de documentos en Internet para su posterior recuperación, lectura y procesamiento por humanos y máquinas, de forma análoga a una base de datos (Berners-Lee, 2006; GmbH Oberpfaffenhofen, 2018).

Los datos ligados, son la forma de publicación en la web semántica, también llamados datos vinculados, datos enlazados (*Linked Data*), o datos abiertos vinculados (*Linked Open Data*) cuando se encuentran disponibles en acceso abierto. Los datos ligados interconectan documentos disponibles en línea principalmente con el uso del esquema *Resource Description Framework* (RDF), el identificador *Uniform Resource Identifier* (URI), el protocolo HTTP, la tecnología de consulta *Protocol and RDF Query Language* (SPARQL) y los enlaces entre sitios web relacionados (Berners-Lee, 2006). Otro de los recursos clave para la publicación de datos ligados son los sistemas de organización del conocimiento, específicamente las ontologías, que representan a los objetos, grupos de objetos y sus relaciones (W3C, 2013; W3C, 2015b).

En ciencias de la vida la web 3.0 es considerada un marco de publicación de datos (Thessen y Patterson, 2011). Mientras que, en biodiversidad, Page (2006) recomienda el uso de tecnologías relacionadas principalmente RDF, para describir conceptos taxonómicos dentro de bases de datos. Amanqui *et al*. (2016) proponen el uso del contexto espacio-temporal para integrar los datos de especies a la web semántica, desde la perspectiva del investigador en biodiversidad. Chawuthai *et al*. (2016) recomiendan el uso de la especificación RDF en recursos de Internet que representan taxones, con el fin de preservar la información asociada a cambios taxonómicos. Baskauf *et al*. (2016) describen una guía para el uso de RDF en el esquema *Darwin Core,* con el fin de relacionar las actividades en informática de la biodiversidad con la publicación de datos ligados y la web semántica.

Datos abiertos

Los datos abiertos (*Open data*), son datos publicados en la *World Wide Web* con derechos de autor y licencias que permiten copiar, acceder y reutilizar libre y gratuitamente la información, al citar debidamente la fuente de procedencia (Berners-Lee, 2006; FOSTER, 2018; NLM, 2019).

Para la ciencia, los datos abiertos tienen la ventaja de acelerar el ritmo de las investigaciones; proporcionan los medios para verificar los resultados y conclusiones; eliminan las barreras técnicas, legales y financieras que fragmentan y aíslan a los datos; y permiten que la información sea compartida, utilizada e intercambiada (Gruen, Houghton y Tooth, 2014). Una de las principales limitaciones para la información de biodiversidad son los datos oscuros o *dark data,* es decir, aquellos datos que no se indexan y almacenan adecuadamente, permanecen ocultos para los usuarios potenciales de la información y finalmente se pierden sin ser aprovechados (Heidorn, 2008; Chavan y Penev, 2011).

Chavan y Penev (2011) revisan el estado actual en el descubrimiento de los datos en biodiversidad, destacan el acceso abierto como un medio para la toma informada de decisiones y proponen el documento de datos como un medio para incentivar la publicación de la información. Costello *et al*. (2014) analizan el acceso abierto a largo plazo en repositorios digitales de especies; mientras que, Groom, Weatherdon y Geijzendorffer (2016) destacan la necesidad de la apertura de datos de fuentes como la ciencia ciudadana. Sikes *et al*. (2016) recomiendan el libre acceso a los datos provenientes de colecciones de historia natural y revisiones taxonómicas, que constituyen un importante conjunto de los datos oscuros en biodiversidad. Davies *et al*. (2017) resaltan la falta de acuerdos para la publicación de datos producto de las investigaciones, lo que representa una barrera para la reproductibilidad de los estudios o el reúso de los datos y metadatos para el análisis de información a grandes escalas o desde otras perspectivas.

Justificación

La información actual sobre la diversidad biológica, es el resultado de importantes inversiones en recursos biológicos, humanos y económicos (Gropp, 2016; Koureas *et al*., 2016). Los datos sobre especies disponibles en la *World Wide Web*, son ampliamente utilizados en investigaciones complejas que analizan o integran la información a grandes escalas geográficas y temporales (Howe *et al*., 2008; Veiga, Cartolano y Saraiva, 2014; Silva *et al*., 2016).

La heterogeneidad inherente a los datos de biodiversidad y la enorme cantidad de repositorios que reúnen información sobre organismos, han ocasionado problemas para el manejo, uso y análisis de grandes conjuntos de datos procedentes de distintas fuentes (Bowker, 2000; Bach *et al*., 2012; Amanqui *et al*., 2016; Koureas *et al.*, 2016). Las limitaciones técnicas, infraestructurales, sociales, políticas, culturales y económicas, implicadas en la publicación de datos, repercuten directamente en el conocimiento y en la conservación de la diversidad de seres vivos; particularmente, las colecciones digitales, reflejan la percepción de la sociedad sobre la biodiversidad (Roberts y Moritz, 2011; Turnhout y Boonman-Berson, 2011; Xiao-Ting, 2012).

Esta tesis, busca contribuir al conocimiento actual sobre la infraestructura disponible para el manejo de datos de especies en la *World Wide Web,* al describir las características de las colecciones digitales de diversidad biológica más relevantes, que reúnen, publican, preservan y movilizan la información global de especies; y al examinar el desarrollo y uso de los recursos electrónicos disponibles para representar, sistematizar, almacenar y distribuir los datos; considerados desafíos para la investigación multidisciplinaria en biodiversidad (Bach *et al*., 2012).

Las soluciones informáticas comunes para el manejo de datos de especies permitirán en un futuro que la información en biodiversidad sea interoperable, se encuentre conectada dentro de la web semántica y pueda ser analizada desde diversas perspectivas, para generar nuevos conocimientos que permitan comprender y preservar la diversidad de la vida (Laurenne *et al*., 2014; Peterson, Soberón y Krishtalka, 2015).

Objetivos

Objetivo general

Describir el estado del conocimiento respecto al desarrollo e implementación de metadatos, estándares y datos ligados en el manejo de datos mundiales de especies en colecciones digitales, a partir de los artículos de investigación y las bases de datos más relevantes.

Objetivos específicos

- Reconocer y caracterizar la literatura científica más destacada sobre el manejo de datos de especies en colecciones digitales disponibles en línea.
- Examinar el panorama actual de las colecciones digitales de diversidad biológica identificadas en los artículos de investigación.
- Evaluar el desarrollo e implementación de metadatos, estándares y datos ligados en el manejo de datos mundiales de especies.
- Crear una base de datos que integre y sistematice los recursos digitales recuperados en esta tesis.

Preguntas de investigación

1. ¿Cuál es el estado en el desarrollo y uso de metadatos, estándares y datos ligados para el manejo de datos mundiales de especies almacenados en colecciones digitales de biodiversidad?
2. ¿Cuáles son las principales características de los artículos de investigación más significativos sobre el manejo de datos de especies?
3. ¿Cuál es el panorama actual de las bases de datos utilizadas en informática de la biodiversidad para almacenar la información?
4. ¿Cuáles son los principales recursos digitales para el manejo de datos y metadatos de biodiversidad reportados en la literatura científica más relevante?
5. ¿Cuáles son las tendencias en la implementación de recursos digitales en las colecciones que almacenan datos de todos los grupos de seres vivos?

Materiales y métodos

El material y método de esta tesis se dividió en cinco etapas que se ilustran en la figura 3 y se detallan en las subsecuentes secciones.

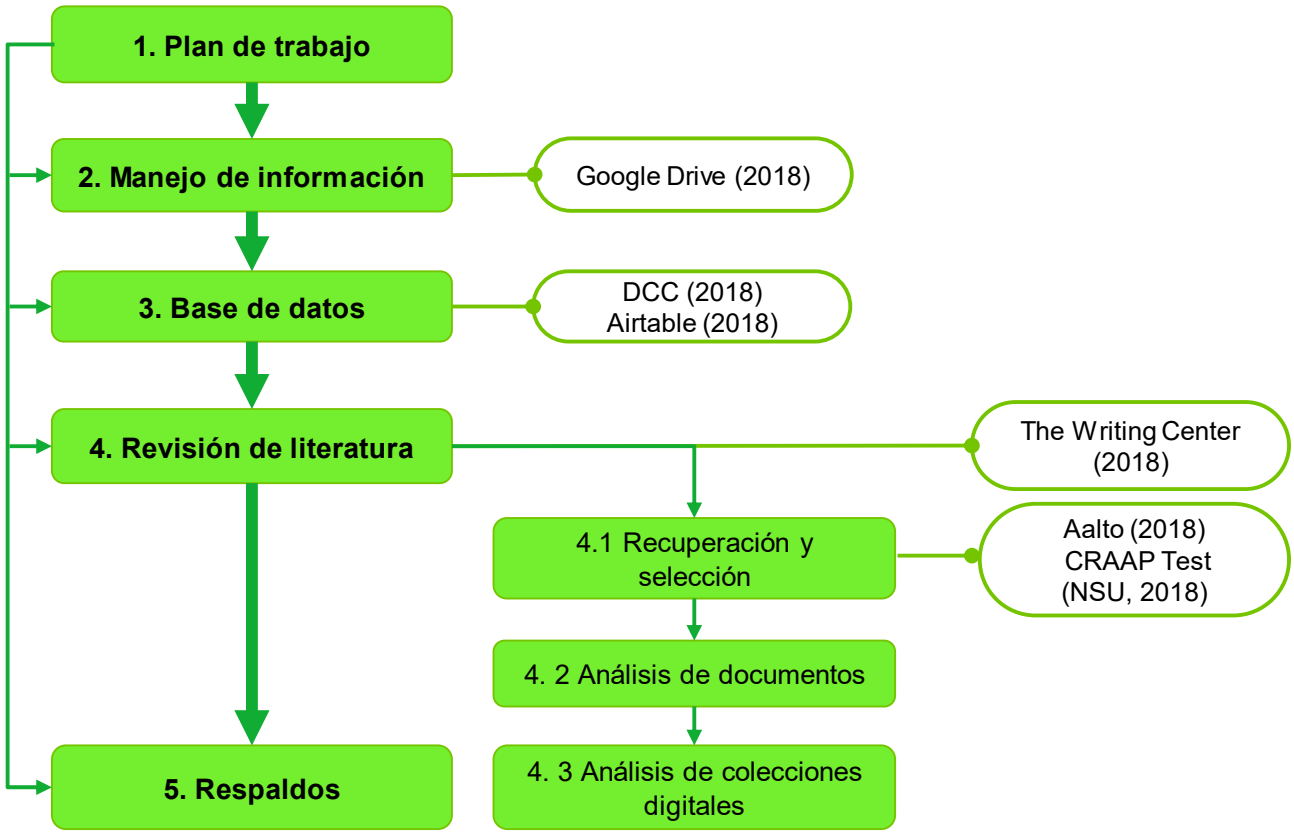


Figura 3. Etapas de trabajo de esta tesis.

1. Plan de trabajo

Las actividades necesarias para cumplir con el objetivo principal de esta tesis se establecieron con base en un horario de 40 horas semanales. La primera etapa consistió en el uso de recursos informáticos para el manejo de información biológica; mientras que, durante la segunda etapa se estableció el problema de investigación, las preguntas, los objetivos y las necesidades de información. El calendario de trabajo determinado puede ser consultado en el anexo 1.

2. Manejo de información

Los protocolos y procesos para el manejo de la información de esta tesis se realizaron a nivel de laboratorio e individual. A nivel de laboratorio, se empleó el sistema de información (SI) creado por la Dra. Layla Michán, que se compone por: la base de datos dentro del manejador *Airtable* (2018); el paquete de soluciones *Google* (2018), para el manejo de documentos y comunicación entre los miembros del laboratorio; y el sistema de organización del conocimiento que establece la estructura del SI. La información a nivel individual, incluyó la creación de las preguntas de investigación, la identificación de las fuentes de información, el diseño de la base de datos del proyecto y el uso de las aplicaciones para automatizar procesos que se describen en la tabla 2.

Tabla 2

Aplicaciones web utilizadas para el manejo de la información de esta tesis

Aplicación web	Uso	Cita
Google Drive	Servicio para la creación, edición y distribución de documentos a nivel laboratorio.	(Google, 2018a)
Inoreader	Lector de *RSS* para recibir actualizaciones de recursos, artículos, consultas y páginas web de relevancia.	(Innologica, 2018)
Chrome (Windows 10/8.1/8/7 64-bit)	Navegador web, utilizado para automatizar tareas con el uso de complementos y marcadores.	(Google, 2018b)
Motor de búsqueda *Google*	Motor de búsqueda para recuperar contenido en Internet.	(Google, 2018c)
Google Hangouts	Servicio de mensajería para la comunicación entre los miembros del laboratorio.	(Google, 2018d)
Inoreader Companion (v. 4.1.5)	Complemento del lector Inoreader para agilizar el registro de notificaciones *RSS* (*Really Simple Syndication*).	(Inoreader, 2018)
Highlight Tool (v. 60)	Complemento para clasificar información dentro de documentos de texto de *Google.*	(Anónimo, 2018)
Advanced URL Shortener (v. 22)	Complemento para reducir el tamaño de direcciones *URL.*	(Digital Thoughts, 2018)
DOAJ API (*v.1.0.0*)	Servicio para realizar consultas en la base de datos *DOAJ.*	(DOAJ, 2018b)
Circos Table Viewer v0.63-9	*Software* para visualizar datos e información en un diseño circular.	(Krzywinski, 2018)

3. Base de datos

La sistematización de los datos recuperados y generados en esta tesis, se realizó con el diseño y creación del repositorio denominado “Colecciones biodiversidad 2018” que forma parte del sistema de información del Laboratorio de bioinformación de la Facultad de Ciencias, UNAM.

La base de datos se fundamentó en el ciclo de curación digital del *The Digital Curation Centre* (DCC, 2018) y los pasos para su creación se resumen en la figura 4. El gestor de base de datos fue el servicio *Airtable* (2018), donde se crearon tablas, registros y campos para almacenar entidades y sus atributos, respectivamente.

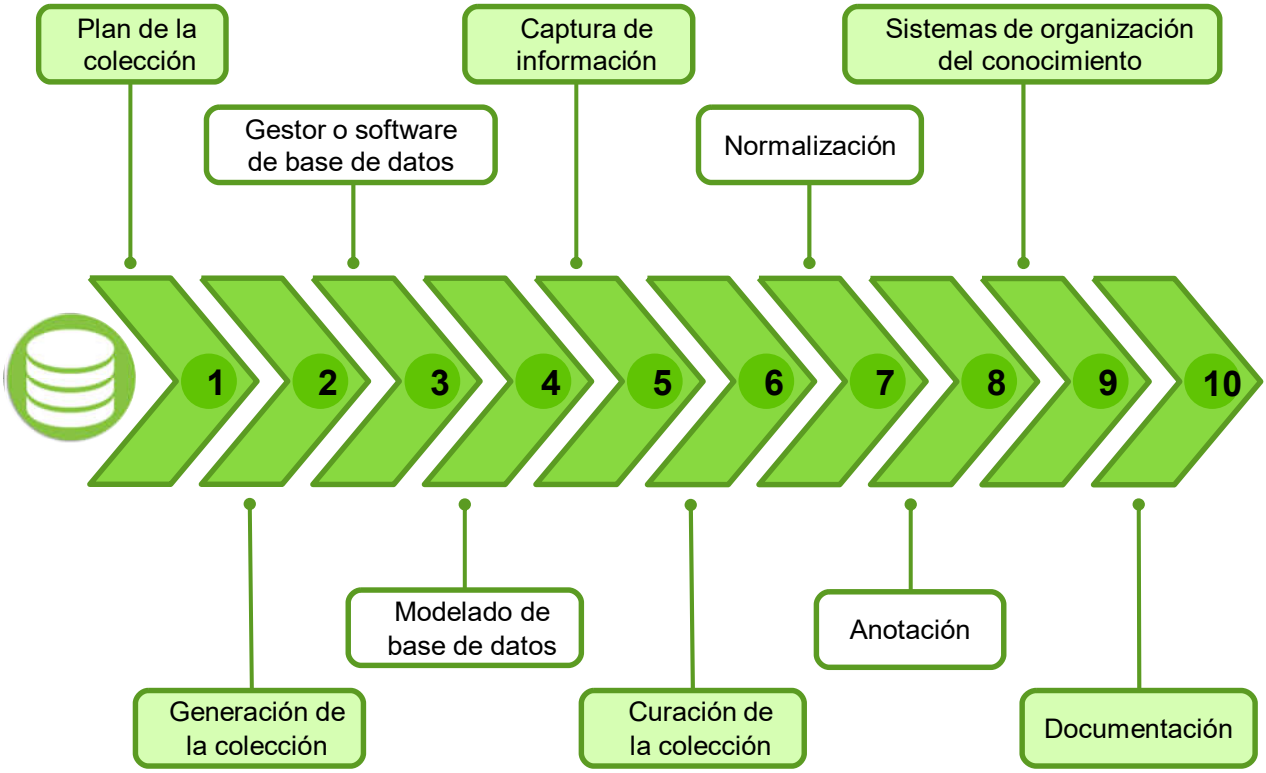


Figura 4. Etapas para la creación de la base de datos “Colecciones biodiversidad 2018”.

La información de la base de datos se añadió a partir de dos procesos: la carga de registros a partir de las colecciones bibliográficas consultadas y la captura manual de objetos digitales de trascendencia para este trabajo. En cada uno de los registros se utilizaron los identificadores de contenido electrónico: *Digital Object Identifier* (DOI) y *Uniform Resource Locator* (URL). La normalización de los datos se realizó con el SOC del laboratorio y los sistemas de organización del conocimiento recuperados para gestionar la información de este trabajo (tabla 3).

Tabla 3

Principales SOC empleados en este trabajo

SOC	Descripción	Cita
Tesauro de la UNESCO	Lista controlada y estructurada de términos para el análisis temático y la búsqueda de documentos en los campos de la educación, cultura, ciencias naturales, ciencias sociales y humanas, comunicación e información.	(UNESCO, 2018)
ID.LOC.GOV	Brinda acceso a los estándares y vocabularios emitidos por la biblioteca del congreso. Incluye los valores de datos y vocabularios controlados que los albergan.	(Library of Congress, 2018)
EDAM Ontology	Ontología integral de conceptos familiares bien establecidos que prevalecen en la bioinformática y la biología computacional, incluidos los tipos de datos e identificadores de datos, formatos de datos, operaciones y temas.	(Ison *et al*., 2013)

Por último, la documentación de la base de datos, incluyó la creación de un diccionario de datos que brinda contexto al contenido de la colección y puede ser consultado en el anexo 2. Los protocolos y procesos necesarios para la interpretación y reúso de los datos se enumeran a continuación:

1. Guía para la recuperación de información de la *Aalto University* (2018).
2. Ciclo de vida de curación digital de la DCC (2018).
3. Test para la evaluación de recursos "CRAAP" (NSU, 2018).
4. Folleto para recuperación de literatura de *The Writing Center* (2018).
5. Normas APA 6a ed. (APA, 2018).

4. Revisión de literatura

Se realizó el análisis de la literatura científica más relevante, de acuerdo con el folleto para revisiones de literatura de *The Writing Center* (2018). El tipo de revisión fue temática, enfocada en las colecciones digitales y el uso de recursos web para el manejo de datos a escala geográfica mundial dentro de la informática de la biodiversidad.

4. 1 Recuperación y selección

La búsqueda y recuperación de los artículos de investigación, se realizó en un proceso de ocho etapas establecidas en referencia al protocolo de recuperación de literatura de la *Aalto University* (2018), como se observa en la figura 5. La explicación detallada de este procedimiento puede ser consultada en el anexo 3.

Figura 5. Fases establecidas para la recuperación de literatura.

Los términos de búsqueda se establecieron en el idioma inglés, a partir de los temas y subtemas de la investigación; el uso de los sistemas de organización del conocimiento descritos en la tabla 3; y la exploración de los términos utilizados por otros autores en documentos de investigación relacionados (tabla 4).

Tabla 4

Términos empleados en la recuperación de literatura

Término español	Término inglés	Sinónimos, abreviaturas o términos relacionados	Enlace
Informática de la biodiversidad	*Biodiversity informatics*	-	https://goo.gl/zN9CzZ
Base de datos	*Database*	*Data banks, collection, DB*	https://goo.gl/mcifpf
Gestión de datos	*Data management*	*Curation*	https://goo.gl/CCtTXt
Dato	*Data*	*Datum, information*	https://goo.gl/corizE
Metadato	*Metadata*	*Cataloguing*	https://goo.gl/Ak8Bdk
Diversidad biológica	*Biological diversity*	*Biodiversity*	https://goo.gl/aX1Yx8
Estándar	*Standard*	-	https://goo.gl/SNnJmf
Datos abiertos	*Open data*	*Open access, Open science*	https://goo.gl/Bab2W3
Datos ligados	*Linked Data*	*Linked Open Data*	https://goo.gl/npQ8ks
Web semántica	*Semantic Web*		https://goo.gl/LSFKHH

Los artículos de investigación se recuperaron a partir de los registros presentes en las bases de datos bibliográficas de mayor alcance en ciencias biológicas: *Scopus* (Elsevier, 2018a), *PubMed* (NCBI, 2018a), *Web of Science* (WoS) [v.5.29], (Clarivate Analytics, 2018a) y el *Directory of Open Access Journals* (DOAJ), (2018a). Las principales características de estas colecciones se presentan en la tabla 5.

Tabla 5

Colecciones bibliográficas utilizadas en la recuperación de literatura

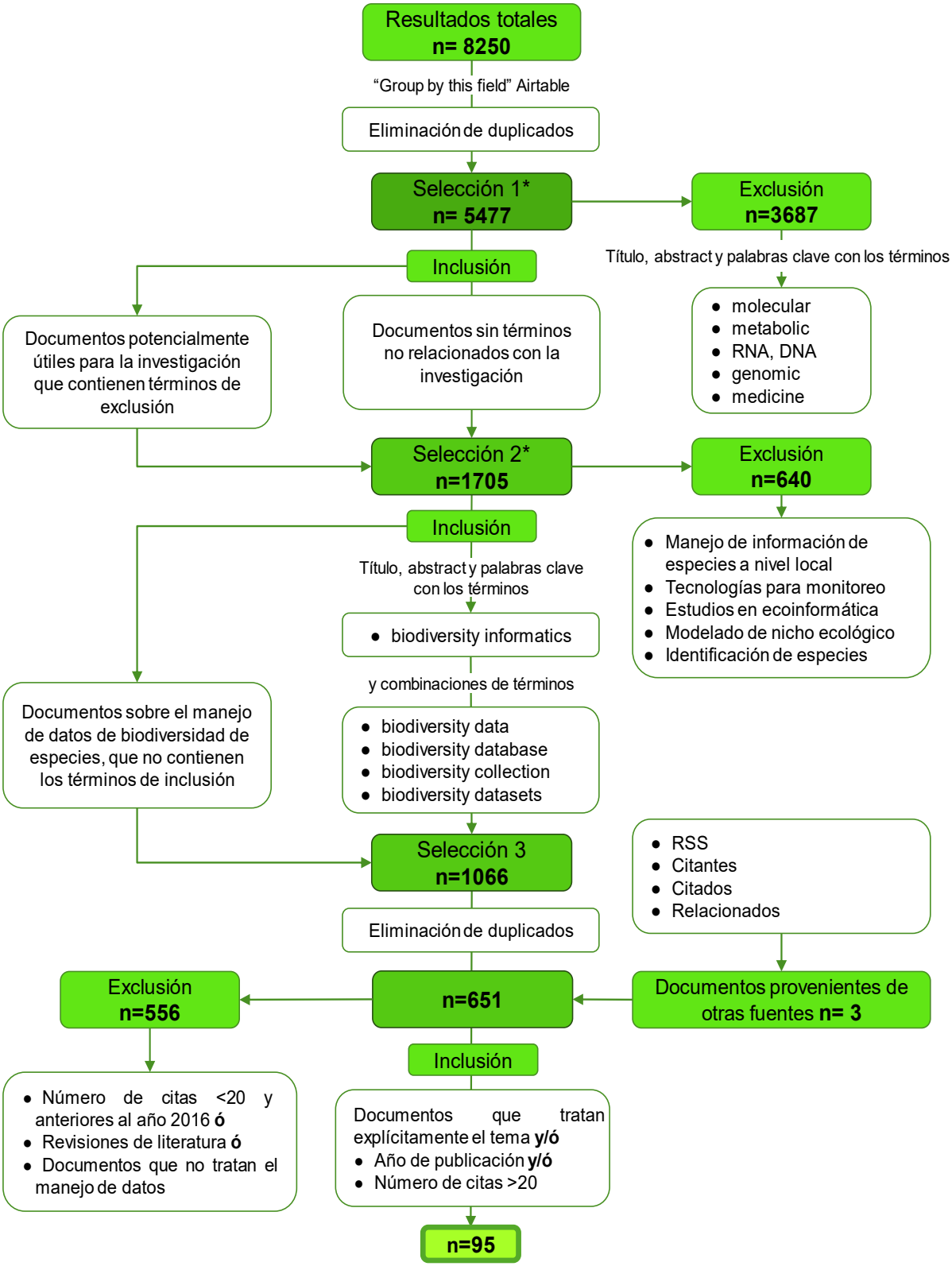
Base de datos	Descripción	Enlace
Scopus	Es la base de datos con el mayor número de resúmenes y citas de literatura revisada por pares. El 15% del contenido de esta base corresponde al área de las ciencias biológicas.	https://www.scopus.com/
PubMed	Comprende más de 28 millones de citas de literatura biomédica de *MEDLINE*, revistas de ciencias de la vida y libros en línea. Las citas pueden incluir enlaces a contenido de texto completo de *PubMed Central* y los sitios web de los editores.	https://www.ncbi.nlm.nih.gov/pubmed/

Tabla 5 (continuación)

Base de datos	Descripción	Enlace
Web of Science [*v*.5.29]	Almacena información de más 3,300 revistas relevantes para las ciencias, ciencias sociales, artes y humanidades, con información publicada desde el año 1900 a la fecha.	https://clarivate.com/products/web-of-science/
Directory of Open Access Journals	Es una colección que indexa y proporciona acceso a revistas y referencias bibliográficas de acceso abierto, revisadas por pares con altos estándares de calidad. Sus servicios y datos se encuentran disponibles de forma gratuita.	https://doaj.org/

Posteriormente, se realizaron consultas en las colecciones bibliográficas, con el uso de los términos de búsqueda y los operadores booleanos “OR”, “AND” y “NOT”. En cada una de las bases de datos, se empleó la guía de recuperación de información, con el fin de utilizar correctamente los campos, filtros y opciones de búsqueda avanzada (NLM, 2016; Clarivate Analytics, 2018b; Elsevier, 2018b; Elasticsearch, 2018). Las consultas finales se descargaron en archivos de texto .csv y .txt separado por tabulaciones y la calidad de la información se certificó con la prueba “CRAAP” de la *Nova Southeastern University* (NSU, 2018), que se detalla en el anexo 4.

Los artículos analizados en esta tesis, se seleccionaron en un proceso de tres etapas, a partir de los registros bibliográficos recuperados. En la primera etapa, se eliminaron los duplicados y artículos no relacionados (figura 6). En la segunda etapa, se identificaron los documentos que tratan explícitamente la gestión de datos digitales de especies, es decir, aquellos que contienen en el título, resumen y/o palabras clave los términos “*biodiversity informatics*”, las combinaciones de términos “*biodiversity data*”, “*biodiversity database*”, “*biodiversity collection*” o “*biodiversity datasets*” y se realizó la identificación manual de investigaciones que no fueron seleccionadas en los filtros previos, de acuerdo a su resumen. Finalmente, en la tercera etapa se agregaron artículos provenientes de notificaciones RSS y citas de otros estudios. Se obtuvo un conjunto de 95 textos científicos relevantes, con base en los siguientes criterios: 1) investigaciones que tienen como objetivo principal el manejo de información mundial de especies en bases de datos y/o 2) documentos publicados a partir del año 2016 y/o 2) documentos con el mayor número de citas publicados antes del año 2016 (>30). El proceso completo de selección se resume en la figura 6.



* Procedimiento realizado en cada una de las bases de datos bibliográficas consultadas

Figura 6. Proceso de selección de literatura.

4. 2 Análisis de artículos

La información presente en los 95 documentos de investigación más relevantes (anexo 5), se analizó detalladamente de forma individual y se registró en la base de datos del proyecto, de acuerdo con las variables descritas en la tabla 6.

Tabla 6

Variables analizadas en la literatura de investigación más relevante

Variable	Descripción
Id revisión	Identificador persistente del artículo (*DOI*, *Handle*, etc.).
Revista	Revista donde fue publicado el documento analizado.
Año	Año de publicación del documento.
Documento	Tipo de documento: artículo, capítulo de libro, conferencia, etc.
Idioma	Lengua en que se encuentra escrito el documento.
Objetivo	Objetivo o problema de estudio tratado en el documento.
Base de datos	Bases de datos citadas o mencionadas en el documento de investigación.
Tipo dato	Tipos de datos de especies dentro de la informática de la biodiversidad.
Metadato	Información proporcionada sobre el uso de metadatos en el manejo de datos de especies.
Esquema de metadatos	Esquemas de metadatos mencionados en el documento de investigación.
Estándar	Información proporcionada sobre el uso de estándares en el manejo de datos de especies.
Tipo estándar	Estándares citados o mencionados en el escrito analizado.
Dato ligado	Información proporcionada sobre el uso de datos ligados en el manejo de la información.
Identificador	Información proporcionada por el autor sobre el uso de identificadores en el manejo de datos de especies.
ID	Identificadores mencionados por los autores para el manejo de datos de biodiversidad de especies.
Acceso abierto	Documento que puede consultarse en acceso abierto.
Dato abierto	Documento que aborda o menciona la implementación de datos abiertos en biodiversidad de especies.

Tabla 6 (continuación)

Variable	Descripción
SOC	Información sobre los sistemas de organización del conocimiento mencionados por los autores para el manejo de datos de biodiversidad de especies.
Vocabulario	Sistemas de organización del conocimiento para el manejo de datos de especies.

Los datos producto del análisis de la literatura, se clasificaron de acuerdo a su tipo en: colecciones digitales, esquemas de metadatos, estándares o especificaciones, sistemas de organización del conocimiento e identificadores. Las colecciones digitales identificadas se verificaron bajo los siguientes criterios: a) alcance geográfico mundial y b) colecciones que almacenan y sistematizan información de biodiversidad a nivel de especies. Por su parte, los recursos digitales se catalogaron de acuerdo a su propósito en: a) recursos de uso general o b) recursos para el manejo de datos de especies y se determinó su frecuencia de aparición en los artículos de investigación.

4. 3. Análisis de colecciones digitales

Las 72 colecciones digitales de especies citadas en los documentos de investigación se clasificaron y analizaron individualmente con las variables especificadas en la tabla 7.

Tabla 7

Variables analizadas en el conjunto total de colecciones digitales de especies

Variable	Descripción
Dominio de información	Dominio de información de la colección de acuerdo con la clasificación de Triebel y colaboradores (2012): datos taxonómico-nomenclaturales (nombres y clasificación), datos de representatividad geográfico-temporal (ocurrencias) y rasgos-datos descriptivos.
Cobertura taxonómica	Lista de los grupos taxonómicos que incluye la base de datos a nivel de dominio, reino, filo o división y/o clase.
Localización geográfica	Localización a nivel de país de la colección digital con base en la dirección IP (*Internet Protocol*)*.

Nota. *Con el uso de la extensión *Open SEO Stats v. 9.6.0.0* (ChromeFans.org, 2017).

Por último, las tendencias en la implementación de metadatos, estándares y datos ligados se examinaron en el subconjunto de 17 bases de datos que indexan datos de especies de todos los grupos de seres vivos, con las variables que se describen en la tabla 8.

Tabla 8

Variables analizadas en las colecciones digitales de seres vivos

Variable	Descripción
Esquemas de metadatos	Esquemas de metadatos utilizados en la colección digital.
Marco de estándares	Conjunto de estándares utilizados en la base de datos.
Identificador	Identificador o identificadores utilizados para ubicar los registros presentes en la base de datos.
SOC	Sistemas de organización del conocimiento utilizados en la base de datos.
Datos ligados	Enlaces entre colecciones digitales de especies.
Acceso	Tipo de acceso a la base de datos y a los registros de la colección.
Formato	Formato de descarga de los datos de la colección.

Con base en los resultados, se representaron las tendencias de la información en tablas y gráficas, y se ilustró la capacidad de intercambio de información entre las colecciones de seres vivos con el uso del programa *Circos Table Viewer* v0.63-9 (Krzywinski, 2018). Las relaciones mostradas en las gráficas de cada aspecto pueden ser consultadas en el anexo 6.

5. Respaldos

Durante todas las etapas de este trabajo se crearon copias de seguridad de los documentos y tablas en distintos medios de almacenamiento. Los respaldos se realizaron de forma parcial al modificar considerablemente algún componente del sistema de información. Los documentos se dividieron de acuerdo a su procedencia en las carpetas *Airtable* y *Google Drive,* mientras que, de acuerdo al tipo de archivo se emplearon las extensiones .docx, .pptx y .xlsx. Por su parte, los respaldos de cada una de las tablas que componen la base de datos se realizaron en archivos de texto con la extensión .csv. El nombre de los archivos se estableció con el nombre del documento original y la fecha de creación del respaldo (ejemplo: 101018anteproyecto.docx).

Resultados y discusión

Como resultado de esta tesis, se identificaron 95 artículos de investigación, 72 colecciones digitales y 131 recursos web para el manejo de datos de especies. Se analizaron 29 variables para dos fuentes de investigación documental: artículos científicos y bases de datos disponibles en línea. La relación entre las fuentes de información, las variables y las secciones de los resultados se resume en la tabla 9. Finalmente, en el último apartado se describe la base de datos “Colecciones biodiversidad 2018” diseñada para integrar la información recuperada.

Tabla 9

Variables y secciones de resultados respecto a su fuente documental

Fuente documental	Variables	Resultado
Artículos de investigación	Revista, año, tipo de documento, idioma	1. Artículos de investigación más relevantes
	Bases de datos, tipos de datos	2. Colecciones digitales de diversidad biológica a nivel de especies
	Metadatos, estándares, SOC, identificadores, datos ligados, datos abiertos	3. Recursos web para el manejo de datos digitales de especies
Bases de datos	Dominio de información, cobertura taxonómica, localización geográfica	2. Colecciones digitales de diversidad biológica a nivel de especies
	Esquemas de metadatos, marco de estándares, identificador, SOC, datos ligados, acceso, formato	4. Implementación de recursos web en colecciones digitales de especies

1. Artículos de investigación más relevantes

Los artículos científicos revisados por pares, son el principal producto de las investigaciones, por lo que son una fuente confiable de información para conocer el estado del conocimiento de una disciplina. Como resultado de la recuperación, evaluación y selección de literatura se obtuvo un conjunto de 95 documentos que contienen información actual, relevante y estrechamente relacionada con el manejo de datos sobre biodiversidad a nivel de especies. Las colecciones bibliográficas donde se indexó el mayor número de documentos fueron *Scopus* (n=77) y *Web of*

Science (n=53). Una gran proporción de documentos científicos son compartidos entre las colecciones bibliográficas consultadas (figura 7). Las colecciones que presentaron el mayor número de documentos de forma individual son *Scopus* (n=21) y *DOAJ* (n=6). La base de datos *DOAJ* fue una de las colecciones bibliográficas más importantes para este trabajo, ya que indexa revistas y artículos en acceso abierto, que representaron contribuciones importantes; como el caso de Bingham *et al. (*2017), quienes realizan uno de los análisis más completos y recientes sobre el panorama actual de la informática de la biodiversidad. Por otro lado, tres artículos de investigación fueron recuperados a partir de otras fuentes, que incluyen notificaciones RSS de las consultas originales, documentos citantes, citados y relacionados.

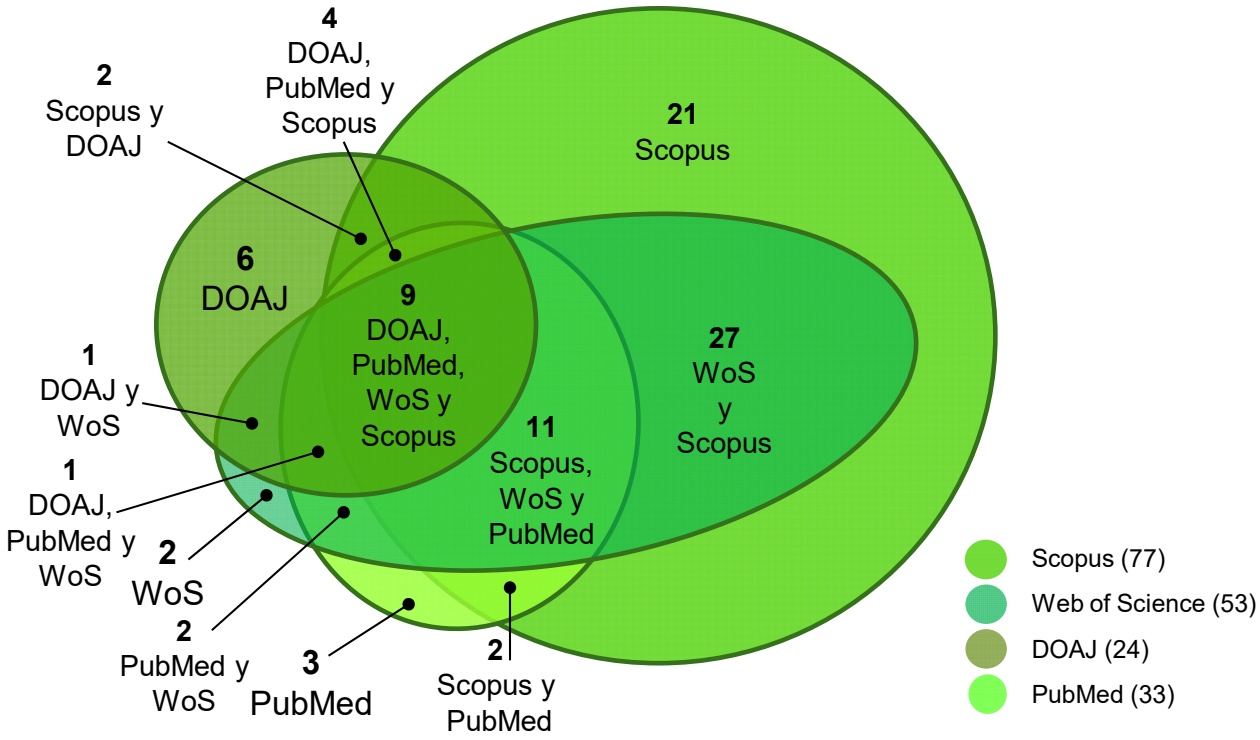


Figura 7. Relación entre el número de artículos de investigación más relevantes y las fuentes bibliográficas consultadas.

Los documentos más significativos identificados en esta tesis, pueden clasificarse en documentos de tipo artículo (n=77), conferencia (n=11), capítulo de libro (n=5) y carta (n=2), tal como se aprecia en la figura 8.

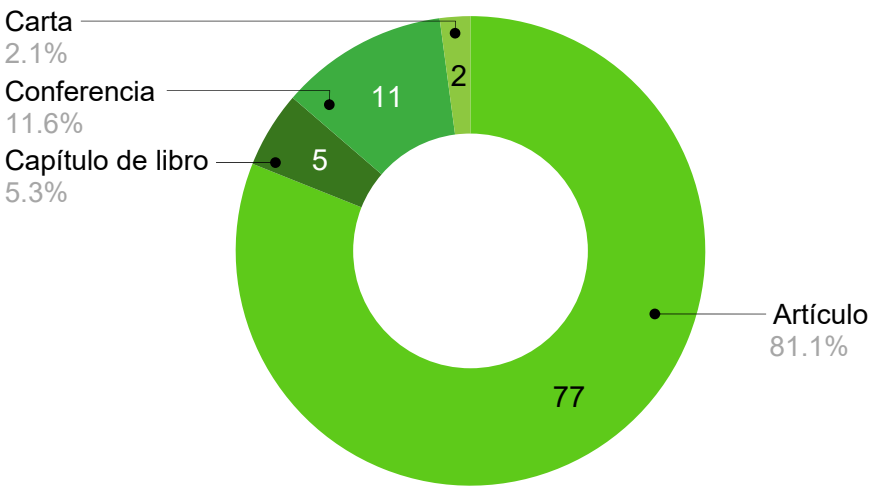


Figura 8. Tipos de artículos científicos más relevantes sobre el manejo de datos de especies.

Los documentos de carácter científico más relevantes en el manejo de datos de especies, fueron publicados en un período de 20 años, a partir de 1998. Entre los años 1998 y 2015 se observa un incremento gradual en el número de artículos más citados, y después del año 2016 los documentos corresponden al criterio de selección de actualidad (figura 9). Se distingue un aumento en la producción científica hasta el año 2017. En el año 2018 se identificaron tres documentos relevantes hasta la fecha en que se realizó el análisis.

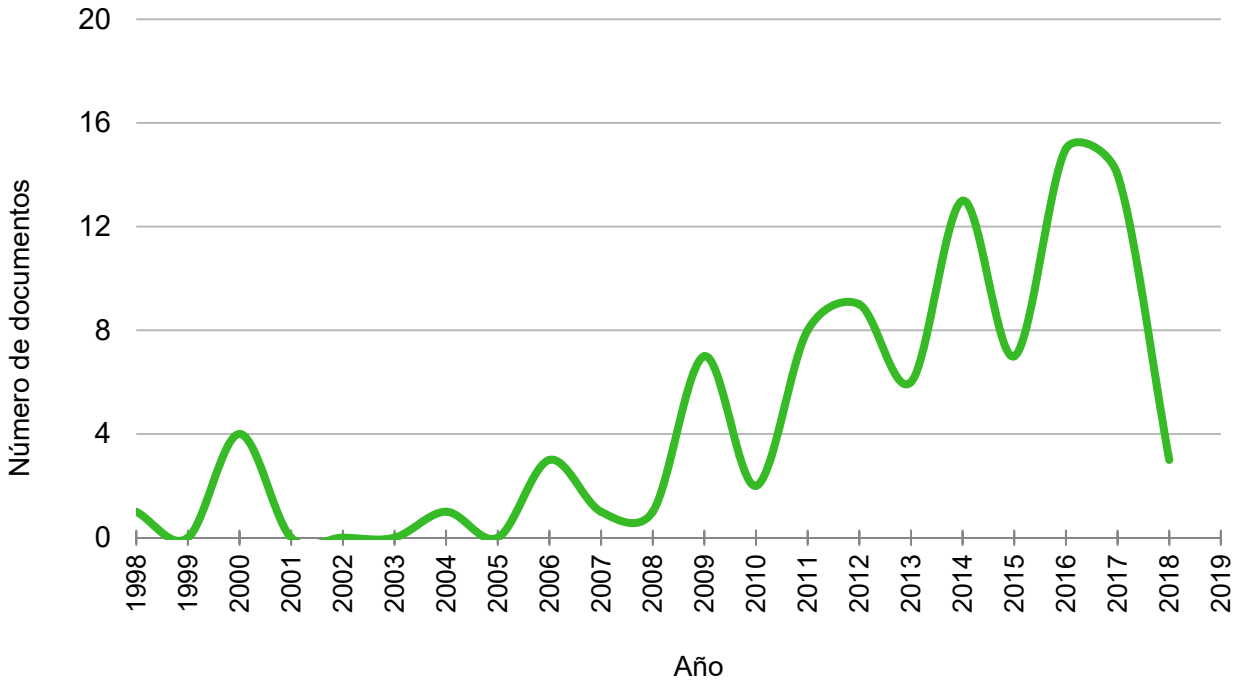


Figura 9. Año de publicación de los principales artículos de investigación identificados.

Los 95 documentos fueron publicados en 59 revistas científicas. Las principales revistas con tres o más artículos de investigación se especifican en la figura 10. El mayor número de artículos pertenece a *PLOS ONE* (n=8) y *BMC Bioinformatics* (n=7), seguidas de *Biodiversity Informatics*, *Semantic Web y Trends in Ecology and Evolution,* con cuatro documentos cada una. Finalmente, *Ecological informatics* y *Biodiversity Data Journal,* presentaron tres documentos de investigación y el resto de artículos se encontró individualmente en 52 revistas. Los documentos que analizan el manejo de datos en informática de la biodiversidad se encuentran publicados en revistas no especializadas en el tema como *MycoKeys* y *Research Ideas and Outcomes (RIO)*.

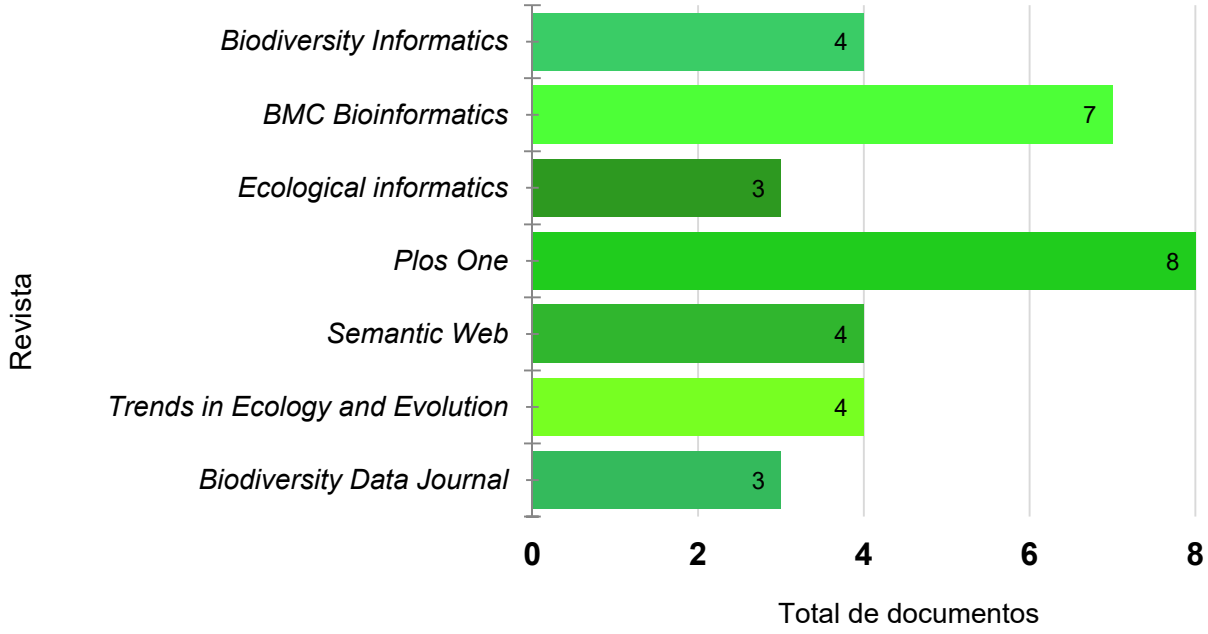


Figura 10. Revistas científicas más destacadas respecto al número de documentos publicados.

El principal idioma de la literatura científica analizada fue el inglés, con 93 artículos; los documentos restantes fueron publicados uno en chino y uno en portugués. El 62% de los artículos (n=59) se encuentra disponible a través de acceso abierto, mientras que, el 39% (n=37) presenta restricciones de suscripción o pago para su consulta. La proporción de textos científicos disponible en acceso abierto, representa una ventaja para la implementación de las recomendaciones e innovaciones tecnológicas, dentro de las colecciones digitales e instituciones encargadas del manejo de datos de especies a distintas escalas.

Se identificó que el 96.84% (n=92) de los documentos, cita o menciona al menos una colección digital de especies. Las investigaciones que enumeran el mayor número de colecciones digitales se ilustran en la figura 11.

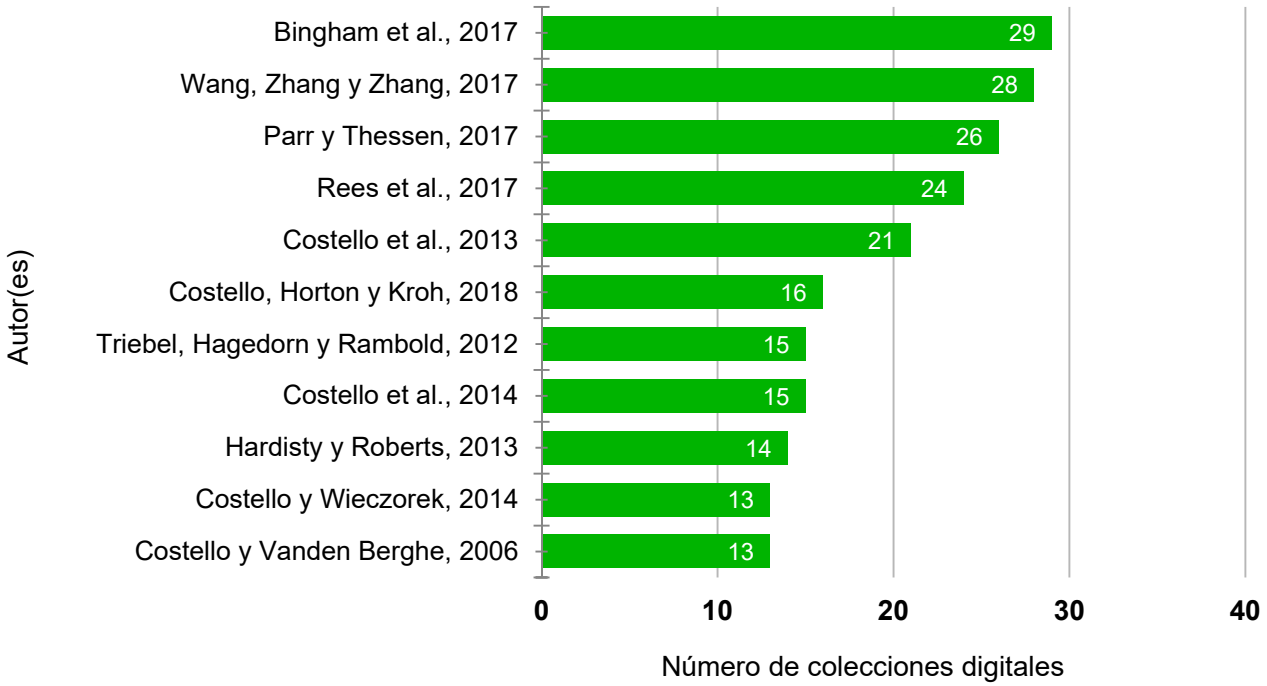


Figura 11. Documentos de investigación más relevantes respecto al número de colecciones digitales identificadas.

Bingham *et al*. (2017) fue la investigación que hace referencia al 40.27% (n=29) de las bases de datos identificadas en esta tesis. Los artículos más relevantes respecto al número de bases de datos, buscan establecer el panorama actual de la informática de la biodiversidad (Bingham *et al*., 2017; Parr y Thessen, 2018), abordan problemas relacionados con la disponibilidad de datos de especies (Wang, Zhang y Zhang, 2017) o bien describen bases de datos especializadas en organismos marinos (Costello *et al*., 2013; Costello, Horton y Kroh, 2018).

Los principales documentos que citan o mencionan el mayor número de recursos web se representan en la figura 12. Estas investigaciones se encuentran relacionadas con la conceptualización de colecciones digitales como Traitbank® (Parr *et al*., 2016) y examinan la gestión de datos en informática de la biodiversidad (Goddard *et al*., 2011; Guralnick *et al*., 2015; Hugo *et al*., 2016).

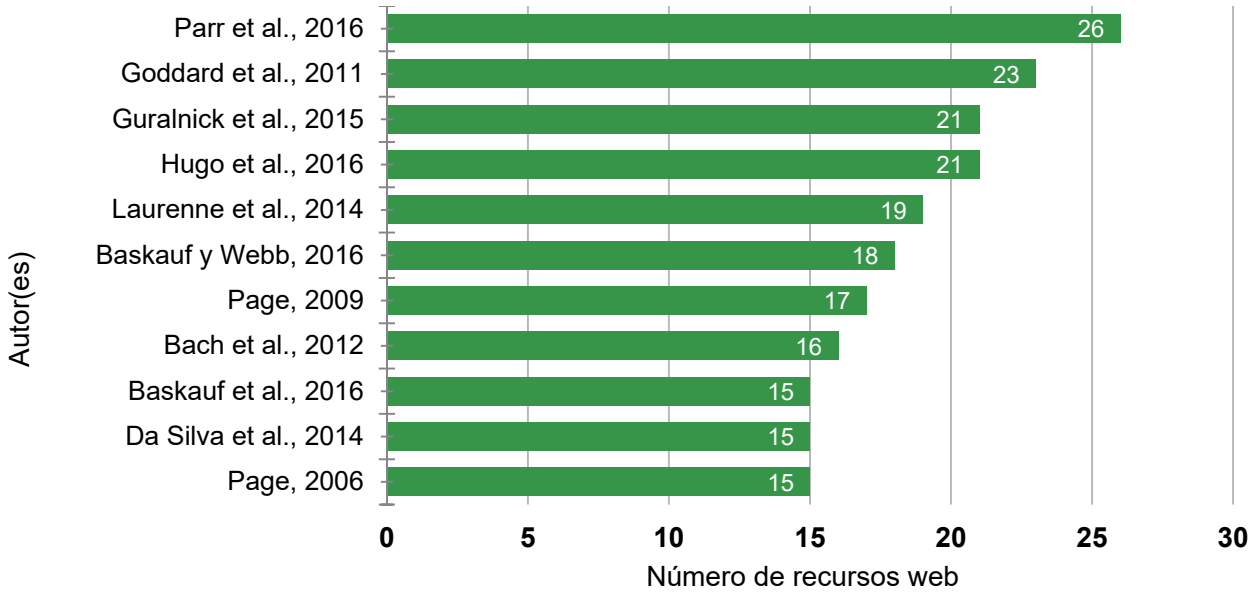


Figura 12. Documentos de investigación más relevantes respecto al número de recursos web mencionados.

El 97.89% (n=93) de los documentos de investigación aborda o menciona el uso de metadatos, estándares y/o datos ligados para la gestión de datos digitales de especies en Internet. Específicamente, 85 documentos mencionan la implementación de estándares, 78 de metadatos y 43 de datos ligados, como se ilustra en la figura 13.

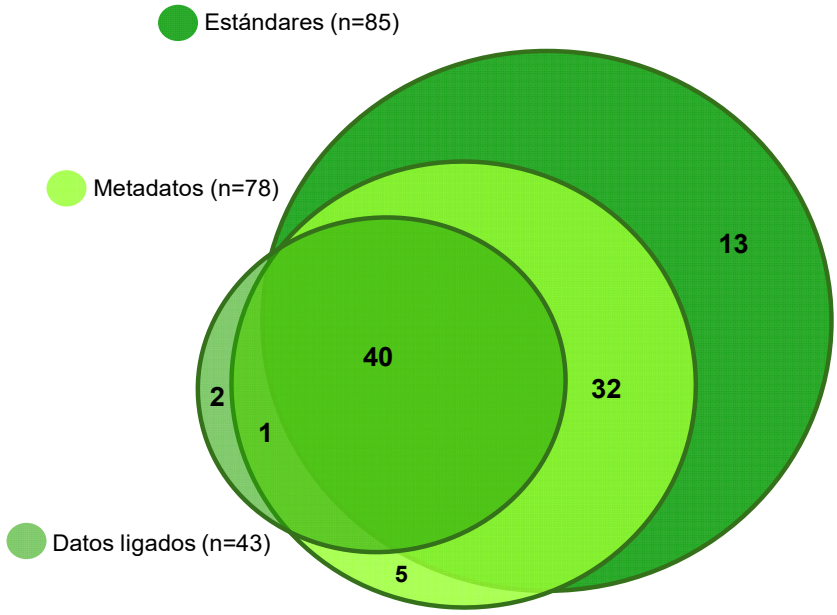


Figura 13. Tipos de recursos web presentes en los artículos de investigación.

Los recursos web, son reconocidos desde las primeras investigaciones como herramientas para el manejo de datos digitales de especies en colecciones. En términos de temas y año de publicación, se observa que los estándares y metadatos son los recursos predominantes en la literatura científica a lo largo del tiempo, mientras que los datos ligados se mencionan por primera vez en el 2006 y han cobrado notoriedad en los últimos años (figura 14). Los datos ligados, son el tipo de recurso menos estudiado, con un menor número de investigaciones en comparación con aquellos estudios que abordan el uso de metadatos y estándares. Los datos ligados son fundamentales para la publicación de datos de especies en la *World Wide Web,* ya que evitan la duplicación de información, relacionan a los datos con otros dominios; y permiten que éstos puedan ser localizados y comprendidos por humanos y máquinas para descubrir patrones en grandes conjuntos de información.

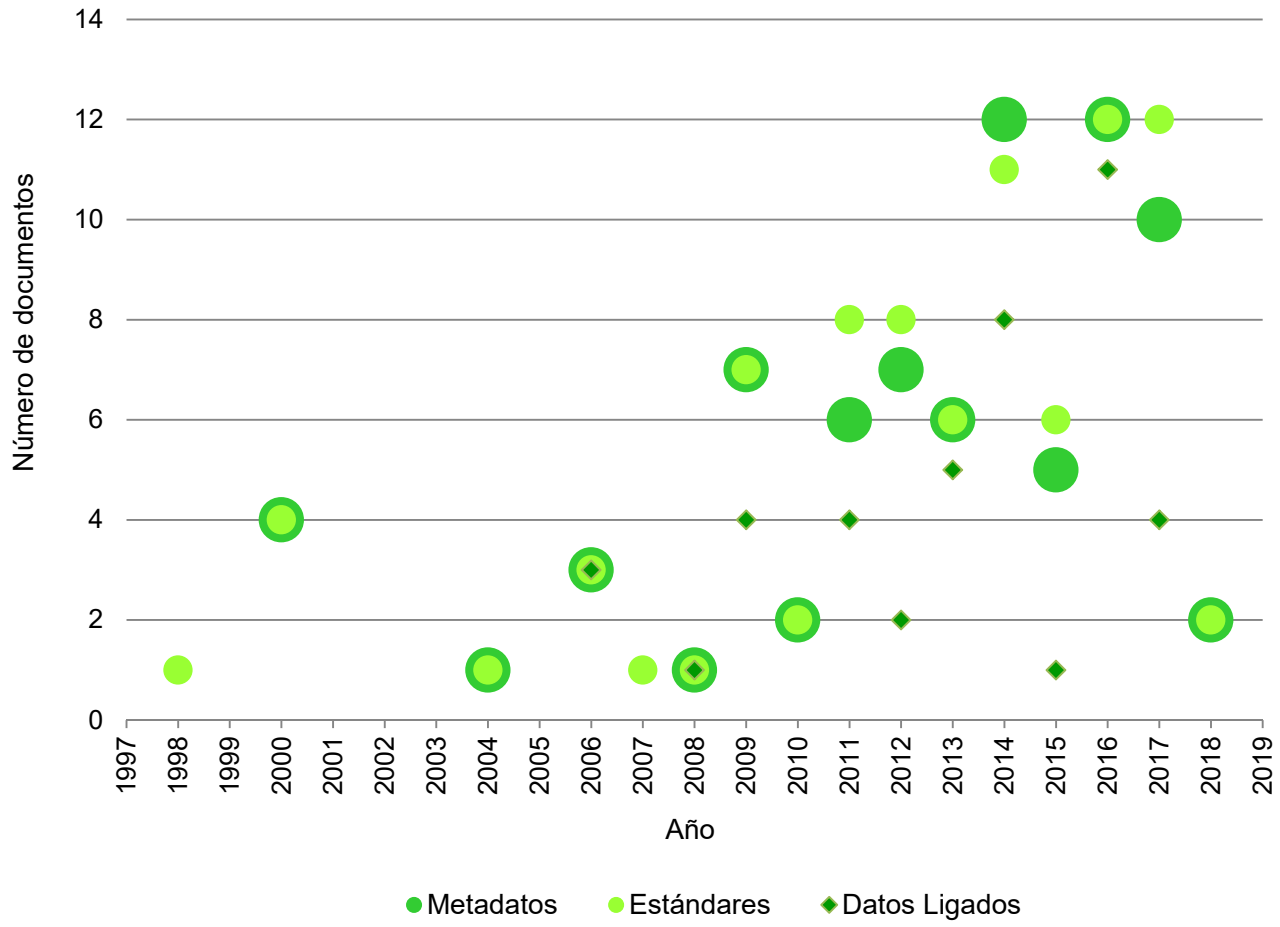


Figura 14. Número y año de publicación de los artículos que examinan el uso de metadatos, estándares y datos ligados.

La producción científica respecto al manejo de datos digitales de especies, a partir de los artículos más relevantes, refleja el creciente interés en el tema en los últimos años. De forma general no se observa una tendencia en la publicación de documentos dentro de revistas especializadas en colecciones o en el manejo de datos digitales, con contribuciones importantes en el tema en revistas como *MycoKeys* y *RIO*.

La mayor proporción de documentos disponibles en acceso abierto, puede reflejar los ideales de la comunidad de la informática de la biodiversidad, sobre la libre disponibilidad de información en Internet. Por esta razón, el uso de bases de datos bibliográficas que indexan revistas y documentos de investigación en acceso abierto es indispensable para la recuperación de investigaciones relevantes.

Las investigaciones descritas en este trabajo no incluyen otras revisiones, que representan la síntesis de la literatura primaria desde otras perspectivas (Bisby, 2000; Edwards, 2000; Guralnick, Hill y Lane, 2007; Page, 2008; Guralnick e Hill, 2009) y pueden resultar de interés para comprender otros aspectos del estado actual de la informática de la biodiversidad.

2. Colecciones digitales de diversidad biológica a nivel de especies

Las bases de datos de carácter global que almacenan información sobre especies en la *World Wide Web*, se identificaron en 92 documentos de investigación que hacen referencia a por lo menos una base de datos. Los repositorios descritos en esta tesis, son iniciativas estables y formalmente reconocidas por los expertos en el manejo de datos, para almacenar, organizar, preservar y distribuir los datos de especies a distintas escalas.

Para comprender el estado actual de las bases de datos de especies, es necesario caracterizar el conjunto de colecciones en línea que componen el panorama general, sin embargo, la falta de listas disponibles, actualizadas y formalmente publicadas, representa una de las principales limitantes. Este trabajo, presenta una contribución al reunir, clasificar y describir 72 colecciones digitales dentro de la informática de la biodiversidad, el mayor número reportado hasta el momento. La lista de repositorios en relación a su cobertura taxonómica y dominio de información se presenta en la tabla 10.

Tabla 10

Colecciones digitales de biodiversidad de especies citadas en los documentos de investigación

Cobertura	Dominio de información	Colecciones digitales
Seres vivos	Taxonomía-nomenclatura	*Catalogue of Life* (http://www.catalogueoflife.org) *Species 2000* (http://www.sp2000.org) *World Register of Marine Species* (http://www.marinespecies.org) *Integrated Taxonomic Information System* (https://www.itis.gov) *Universal Biological Indexer and Organizer* (http://www.ubio.org) *TreeBASE* (https://treebase.org) *NCBI Taxonomy* (https://www.ncbi.nlm.nih.gov/taxonomy) *Open Tree of Life* (https://tree.opentreeoflife.org) *Plazi* (http://www.plazi.org) *The Interim Register of Marine and Nonmarine Genera* (http://www.irmng.org)
	Geográfico-temporal	*Global Biodiversity Information Facility* (https://www.gbif.org) *Ocean Biogeographic Information Systems* (http://www.iobis.org) *Integrated Digitized Biocollections* (https://www.idigbio.org) *Arctos* (https://arctosdb.org)
	Rasgos-datos descriptivos	*Encyclopedia of Life* (http://eol.org) *Global Invasive Species Database* (http://www.iucngisd.org) *Global Invasive Species Information Network* (http://www.gisin.org) *TraitBank* (http://eol.org/info/516) *Global Register of Introduced and Invasive Species* (http://www.griis.org)

Tabla 10 (continuación)

Cobertura	Dominio de información	Colecciones digitales
Animales	Taxonomía-nomenclatura	*ZooBank* (http://www.zoobank.org) *Hexacorallians of the World* (http://hercules.kgs.ku.edu/hexacoral/anemone2/index.cfm) *Avibase* (https://avibase.bsc-eoc.org) *BioNames* (http://bionames.org) *World Database of Free-Living Marine Nematodes* (http://nemys.ugent.be) *Amphibian Species of the World* (http://research.amnh.org/vz/herpetology/amphibia) *Mammal Species of the World* (http://www.departments.bucknell.edu/biology/resources/msw3) *Freshwater Animal Diversity Assessment* (http://fada.biodiversity.be) *The Reptile Database* (http://reptile-database.reptarium.cz) *Turbellarian Taxonomic Database* (http://turbellaria.umaine.edu) *Nomenclator Zoologicus* (http://ubio.org/NomenclatorZoologicus)
	Geográfico-temporal	*VertNet* (http://vertnet.org) *eBird* (https://www.ebird.org/home) *Ocean Biogeographic Information System Spatial Ecological Analysis of Megavertebrate Populations* (http://seamap.env.duke.edu) *Global Assessment of Reptile Distributions* (http://www.gardinitiative.org) *Movebank* (https://www.movebank.org)
	Rasgos-datos descriptivos	*FishBase* (http://www.fishbase.org) *AmphibiaWeb* (https://amphibiaweb.org) *AntWeb* (https://www.antweb.org) *Macaulay Library* (https://www.macaulaylibrary.org) *CephBase* (http://cephbase.eol.org) *Cetabase* (http://cetaceos.webs.ull.es/cetabase.info) *Phenoscape* (https://phenoscape.github.io) *State of the World's Sea Turtles* (http://seamap.env.duke.edu/swot) *Global Ants Database* (http://globalants.org) *Species360* (https://www.species360.org)

Tabla 10 (continuación)

Cobertura	Dominio de información	Colecciones digitales
Plantas	Taxonomía-nomenclatura	*The International Plant Names Index* (http://www.ipni.org) *The Plant List* (http://www.theplantlist.org) *International Legume Database & Information Service* (http://www.ildis.org) *World Flora Online* (http://www.worldfloraonline.org) *Plants of the World Online* (http://www.e-monocot.org)
	Geográfico-temporal	*GlobalTreeSearch* (http://www.bgci.org/global_tree_search.php)
	Rasgos-datos descriptivos	*TRY Plant Trait Database* (https://www.try-db.org/TryWeb/Home.php) *Global Plants on JSTOR* (http://plants.jstor.org) *Botanical Information and Ecology Network* (https://bien.nceas.ucsb.edu/bien/biendata) *Global Inventory of Floras and Traits* (http://gift.uni-goettingen.de)
Hongos	Taxonomía-nomenclatura	*MycoBank* (http://www.mycobank.org) *Species Fungorum* (http://www.speciesfungorum.org)
	Rasgos-datos descriptivos	*A Global Information System for Lichenized and Non-Lichenized Ascomycetes* (http://www.lias.net)
Algas	Taxonomía-nomenclatura	*Algaebase* (http://www.algaebase.org)
Bacterias	Rasgos-datos descriptivos	*The Bacterial Diversity Metabase* (https://bacdive.dsmz.de)
Virus	Taxonomía-nomenclatura	*ICV Taxonomy* (https://talk.ictvonline.org/taxonomy)
Varios*	Taxonomía-nomenclatura	*Index Fungorum* (http://www.indexfungorum.org) *Index to Organism Names* (http://www.organismnames.com) *Tropicos* (http://www.tropicos.org) *Fossilworks* (http://fossilworks.org)
	Geográfico-temporal	*iNaturalist* (https://www.inaturalist.org) *The Paleobiology Database* (https://paleobiodb.org)
	Rasgos-datos descriptivos	*International Union for Conservation of Nature and Natural Resources* (http://www.iucnredlist.org) *SeaLifeBase* (http://sealifebase.org) *Predicts* (https://www.predicts.org.uk/) *Convention on International Trade in Endangered Species of Wild Fauna and Flora Trade Database* (https://trade.cites.org) *Species+* (http://www.speciesplus.net)

Fuente: Elaboración propia.

Las colecciones digitales más importantes, de acuerdo a su frecuencia de aparición en la literatura se presentan en la figura 15. La colección citada en el 83% (n=77) de los documentos de investigación es la *Global Biodiversity Information Facility* (GBIF), que indexa información sobre la representatividad geográfico-temporal de las especies, seguida por *Catalogue of Life* (COL) que reúne los nombres, relaciones y distribución de más de 1.8 millones de especies, dentro del dominio de datos taxonómico-nomenclatural (COL, 2018).

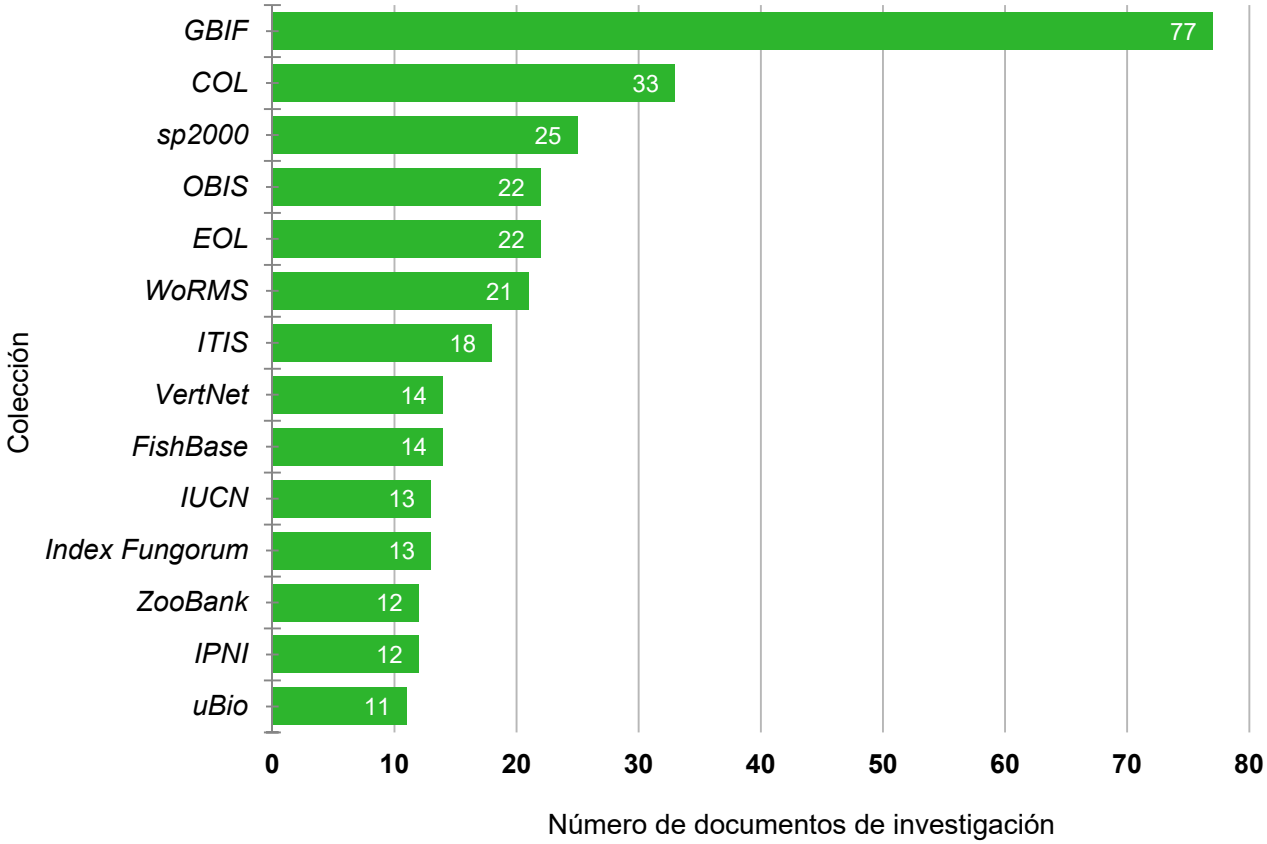


Figura 15. Colecciones digitales más relevantes respecto al número de citas en los documentos de investigación.

2. 1 Tipos de datos en colecciones digitales de especies

Un aspecto fundamental para comprender el panorama actual del manejo de datos de especies disponibles en la *World Wide Web*, es conocer el tipo de información almacenada en las colecciones de biodiversidad de carácter global. Los tipos de datos, las actividades y soportes descritos en las principales investigaciones, se representan en la figura 16 y pueden consultarse detalladamente en el anexo 7.

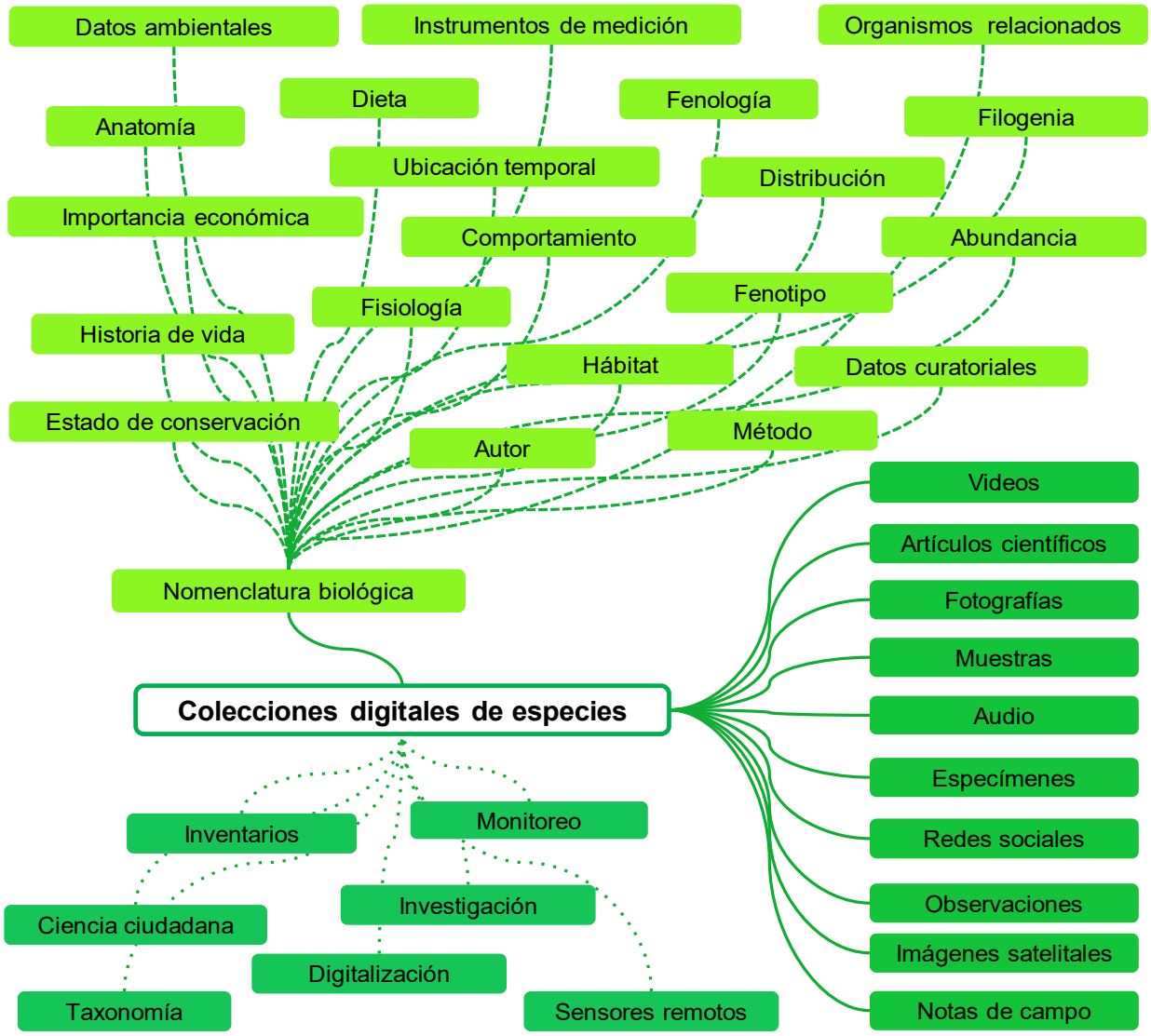


Figura 16. Información almacenada en colecciones digitales de biodiversidad a nivel de especies (Elaboración propia).

Los datos y metadatos que documentan especies son sumamente heterogéneos, ya que provienen de actividades con una gran variedad de propósitos (Bowker, 2000; Bach *et al*., 2012; Walls *et al*., 2014). Esta heterogeneidad, representa una de las principales limitantes y desafíos para el descubrimiento y la calidad de la información en biodiversidad (Wieczorek *et al*., 2012).

Entre las actividades que generan datos digitales de especies podemos encontrar principalmente a las investigaciones científicas (Bach *et al*., 2012), la taxonomía (Martellos y Attorre, 2012), el uso de nuevas tecnologías como los sensores remotos (Juffe-Bignoli *et al*.,

2016), la digitalización de colecciones biológicas (Page, 2013; Beckstein *et al*., 2014), el monitoreo e inventario de organismos (Bach *et al*., 2012; Groom, Weatherdon y Geijzendorffer, 2016) y la ciencia ciudadana (Beckstein *et al*., 2014; Groom, Weatherdon y Geijzendorffer, 2016; Juffe-Bignoli *et al*., 2016). La ciencia ciudadana, es una actividad que ha cobrado notoriedad reciente, al proveer una extensa cobertura temporal y espacial de datos de observación de organismos (Ruete, 2015).

La información de especies, se encuentra almacenada en distintos formatos y soportes como videos (Goddard *et al*., 2011; Costello y Wieczorek, 2014; Baskauf *et al*., 2016), artículos científicos (Cotter y Bauldock, 2000; Page, 2009), fotografías (Goddard *et al*., 2011; Beckstein *et al*., 2014; Costello y Wieczorek, 2014; Sarr *et al*., 2014), muestras (Cotter y Bauldock, 2000; Page, 2009), grabaciones de audio (Cotter y Bauldock, 2000; Sarr *et al*., 2014), especímenes (Martellos y Attorre, 2012; Page, 2013; Santos y Branco, 2012; Costello *et al*., 2014; Peterson, Soberón y Krishtalka, 2015; Guenard *et al*., 2017; Suhrbier *et al*., 2017), redes sociales (Beckstein *et al*., 2014), imágenes satelitales (Cotter y Bauldock, 2000), notas de campo (Goddard *et al*., 2011) y observaciones (Goddard *et al*., 2011; Costello y Wieczorek, 2014; Costello *et al.*, 2014). Los datos que provienen únicamente de observaciones, tienen la desventaja de representar objetos y eventos que no pueden ser verificados; no obstante, es importante aclarar que toda la información de especies, es producto de observaciones e interpretaciones humanas que han sido registradas (Costello *et al*., 2014).

Como se observa en la figura 16, los metadatos asociados a organismos incluyen información sobre los creadores, los métodos o instrumentos empleados en la toma de datos; además de metadatos curatoriales, que no describen propiamente la biología del organismo, pero permiten evaluar la calidad de la información de acuerdo a su procedencia (Bowker, 2000; Costello, 2009). Por otro lado, la información espacio-temporal posibilita reunir datos que han sido colectados por diferentes investigadores o registradores a distintas escalas espaciales (Edwards, Lane y Nielsen, 2000). Aspectos como la importancia económica, o estado de conservación, facilitan la agrupación de información de especies en colecciones con fines prácticos, económicos, políticos, etc. Debido a que los datos digitales de especies, son manejados en sistemas informáticos, se debe reconocer la importancia de los metadatos de carácter técnico que precisan las herramientas con las que han sido manipulados los datos a lo largo de su ciclo de vida.

2. 2 Cobertura taxonómica

El total de colecciones digitales de especies identificadas en esta tesis, puede clasificarse en ocho categorías considerando su cobertura taxonómica. Los grupos identificados fueron: 1) seres vivos, con bases de datos que indexan información de todos los grupos de organismos; 2) animales; 3) plantas; 4) hongos; 5) algas; 6) bacterias; 7) virus; y 8) varios, donde se reúnen repositorios que integran datos de dos o más grupos de seres vivos, pero no incluyen a todos los grupos conocidos (figura 17).

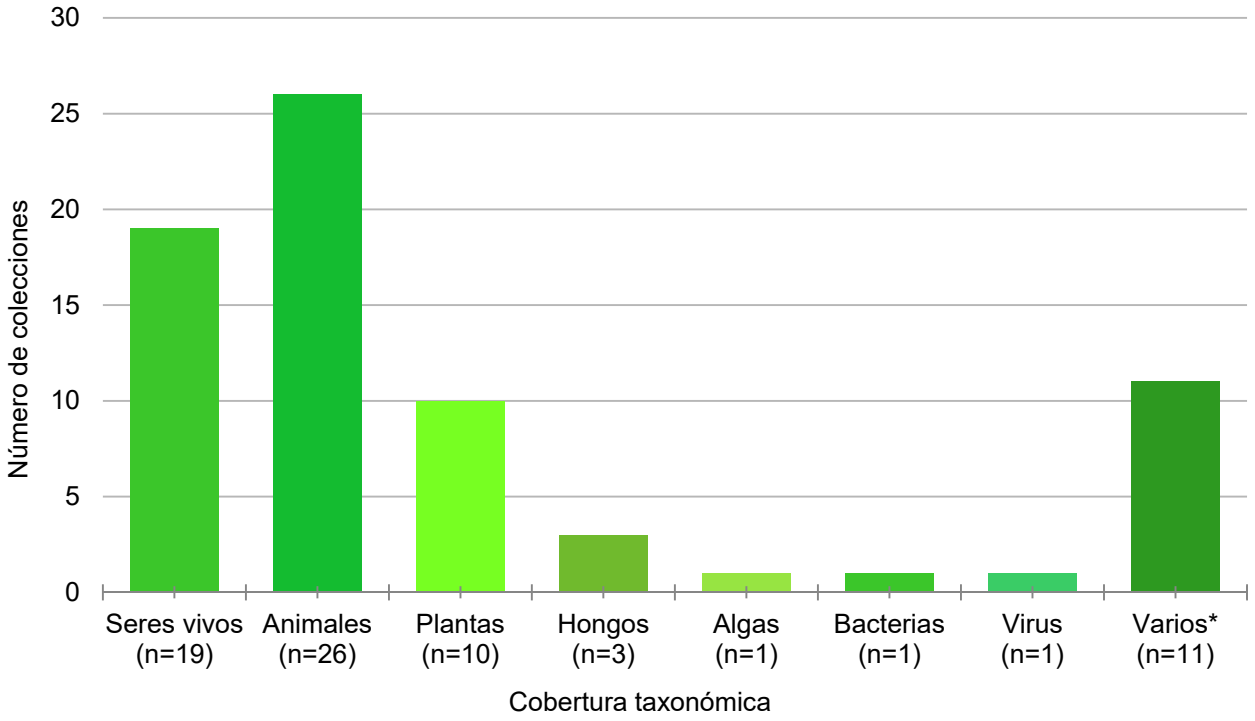


Figura 17. Cobertura taxonómica de las colecciones digitales de especies más relevantes.

El mayor número de colecciones digitales identificado en los documentos de investigación pertenece al grupo de los animales (figura 17); dentro de este conjunto destacan las bases de datos *VertNet* (2018) y *ZooBank* (ICZN, 2018), citadas por una cantidad importante de autores. El siguiente grupo más relevante es el de las colecciones de seres vivos, donde sobresalen los repositorios *GBIF* (GBIF.org, 2018b), *Catalogue of Life* (COL, 2018), *Species2000* (2018), *Encyclopedia of Life* (EOL, 2018a) e *Integrated Taxonomic Information System* (ITIS, 2017). El tercer conjunto más significativo, fue el de las bases de datos que reúnen registros de varios tipos de organismos, sin incluir a todos los grupos; estas colecciones indexan datos con fines prácticos o con importancia económica o política como la *International Union for Conservation of*

Nature and Natural Resources (IUCN, 2018a) y la *Convention on International Trade in Endangered Species of Wild Fauna and Flora Trade Database* (UN Environment, 2018).

De forma general, la cobertura taxonómica se relaciona con aquellos grupos de organismos mejor estudiados como las plantas y los animales. Dentro del grupo de los animales, el 53.84% corresponde a colecciones digitales de vertebrados, mientras que en las plantas el 70% de los repositorios indexa datos de todos los grupos; lo que puede implicar que las bases de datos generales se encuentren en menor número, con respecto a aquellas altamente especializadas. Las colecciones de grupos como los hongos, las algas, las bacterias y los virus se encuentran mal representadas en los documentos de investigación sobre el manejo de datos de especies. Finalmente, la cobertura y similitud de las colecciones digitales identificadas en esta tesis, nos indica que probablemente existe duplicación de esfuerzos, como ha sido reportado por Bingham *et al*. (2017).

2. 3 Dominios de información

De acuerdo con los dominios de información propuestos por Triebel, Hagedorn y Rambold (2012) la mayor proporción de colecciones digitales se ubicó dentro del dominio de datos de taxonomía-nomenclatura, con un 47.2% (n=34), seguido del dominio de rasgos-datos descriptivos con un 36.1% (n=26) y del dominio de representación geográfico-temporal 16.7% (n=12), (ver figura 18).

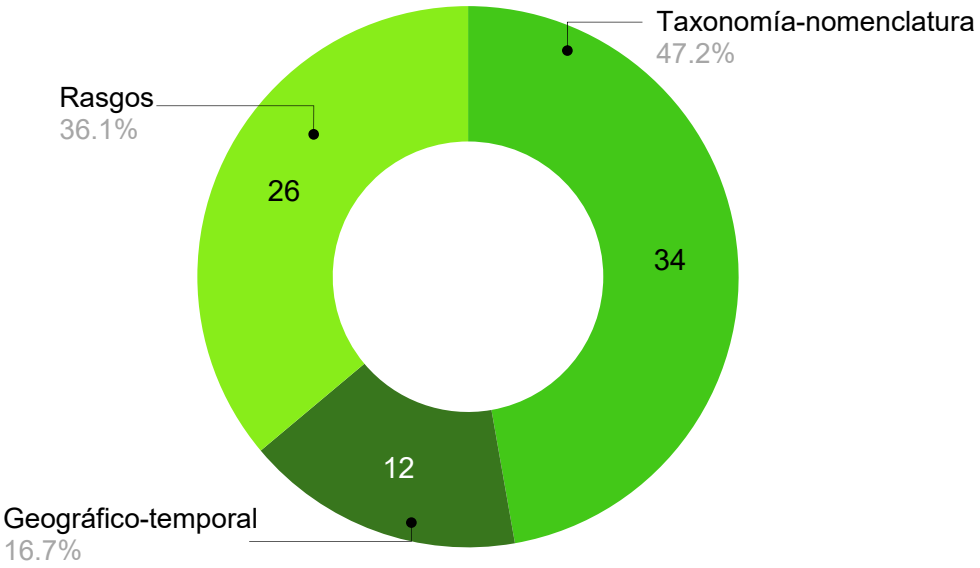


Figura 18. Dominios de información cubiertos por las colecciones digitales de especies.

Estos resultados, muestran que el mayor número de colecciones digitales indexa datos del dominio de información de rasgos-datos descriptivos, que constituyen la información elemental en la toma de datos de biodiversidad. Sin embargo, se observó que los artículos de investigación sobre el manejo de datos de especies, citan o mencionan con mayor frecuencia colecciones particulares que pertenecen a los dominios geográfico-temporal y taxonómico-nomenclatural.

2. 4 Distribución mundial

La distribución mundial de las colecciones digitales de biodiversidad, respecto a su ubicación IP, nos muestra que las 72 bases de datos identificadas en esta tesis se localizan en Estados Unidos y Europa (figura 19). Específicamente el 51% (n=37) de los repositorios se ubica en Estados Unidos, mientras que el 43% (n=31) se concentra en los países de la Unión Europea y Suiza. El resto de las bases de datos, que corresponden al 6.9% (n=5) del total, se ubican en Brasil, Australia y Canadá (figura 19).

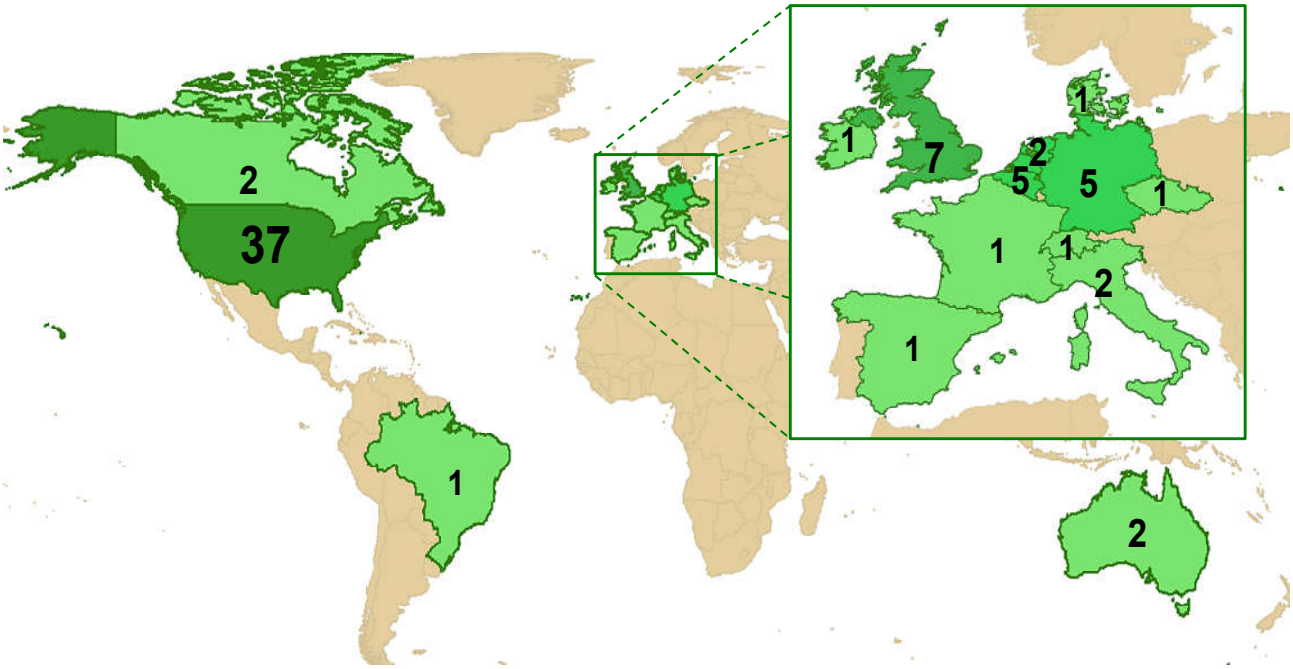


Figura 19. Distribución mundial de las colecciones digitales de diversidad biológica.

La localización geográfica de las colecciones digitales de diversidad biológica pone en evidencia el desarrollo actual de las actividades relacionadas con la gestión de datos mundiales.

Refleja las tendencias en el desarrollo de la infraestructura social y técnica para la publicación y movilización de los conjuntos de datos de especies en la *World Wide Web*, sin desestimar, que parte de esta infraestructura se localiza a escalas regionales y locales no consideradas para este análisis.

La distribución actual de las colecciones digitales de especies, responde a la disponibilidad desigual de la información en biodiversidad, puesto que, la mayor cantidad de datos se concentra en los países desarrollados (Edwards, Lane, y Nielsen, 2000; Canhos *et al.*, 2004; Chavan e Ingwersen, 2009). Se reconoce también, que los desarrollos tecnológicos y actividades de investigación en informática de la biodiversidad, se restringen a algunas instituciones alrededor del mundo (Canhos *et al*., 2004). La disponibilidad de recursos económicos para recopilar y mantener grandes conjuntos de datos de especies, es un factor determinante, ya que, en organizaciones de investigación y gestión de datos de algunas regiones del mundo como Europa, se invierten importantes cantidades de recursos en el desarrollo de herramientas para el manejo de la información (Hugo *et al*., 2016; Koureas *et al*., 2016). Los movimientos recientes de la ciencia abierta, han contribuido a su vez en la unión de algunos países para movilizar datos y promover la colaboración científica en biodiversidad (Costello *et al*., 2013).

Las bases de datos de especies en línea, representan nuestro conocimiento actual libremente disponible sobre todos los organismos conocidos del planeta, y son una opción práctica para reunir los datos mínimos que permitan preservar la biodiversidad (Bowker, 2000).

3. Recursos web para el manejo de datos digitales de especies

Como resultado del análisis de los principales documentos de investigación, se identificaron 131 recursos web para la gestión de datos y metadatos de organismos, presentes en 74 investigaciones. De estos 131 recursos digitales, el 12.97% (n=17) pertenece a esquemas de metadatos, el 38.93% (n=51) a estándares, lenguajes y protocolos, el 29% (n=38) a sistemas de organización del conocimiento y el 19% (n=25) a identificadores.

El 31.57% (n=47) de los recursos web identificados son específicos para el manejo de datos de especies; tal como se observa en la figura 20, pertenecen en su mayoría a sistemas de organización del conocimiento, con un 44.7% (n=21) y a esquemas de metadatos con el 29.8% (n=14).

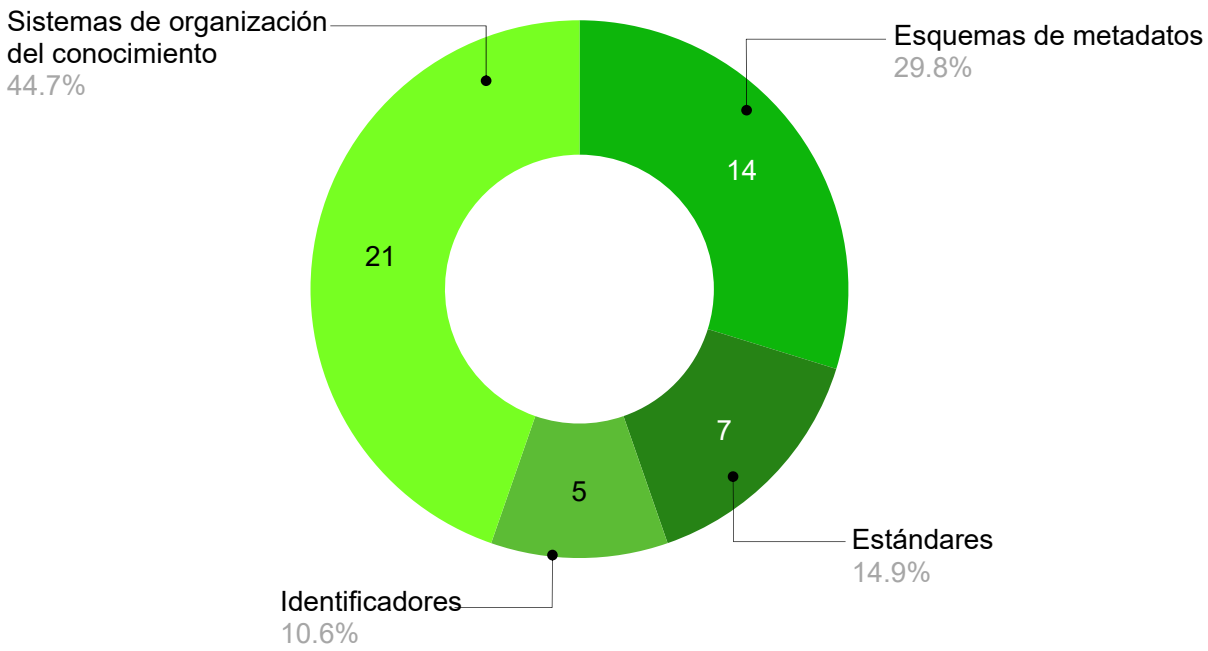
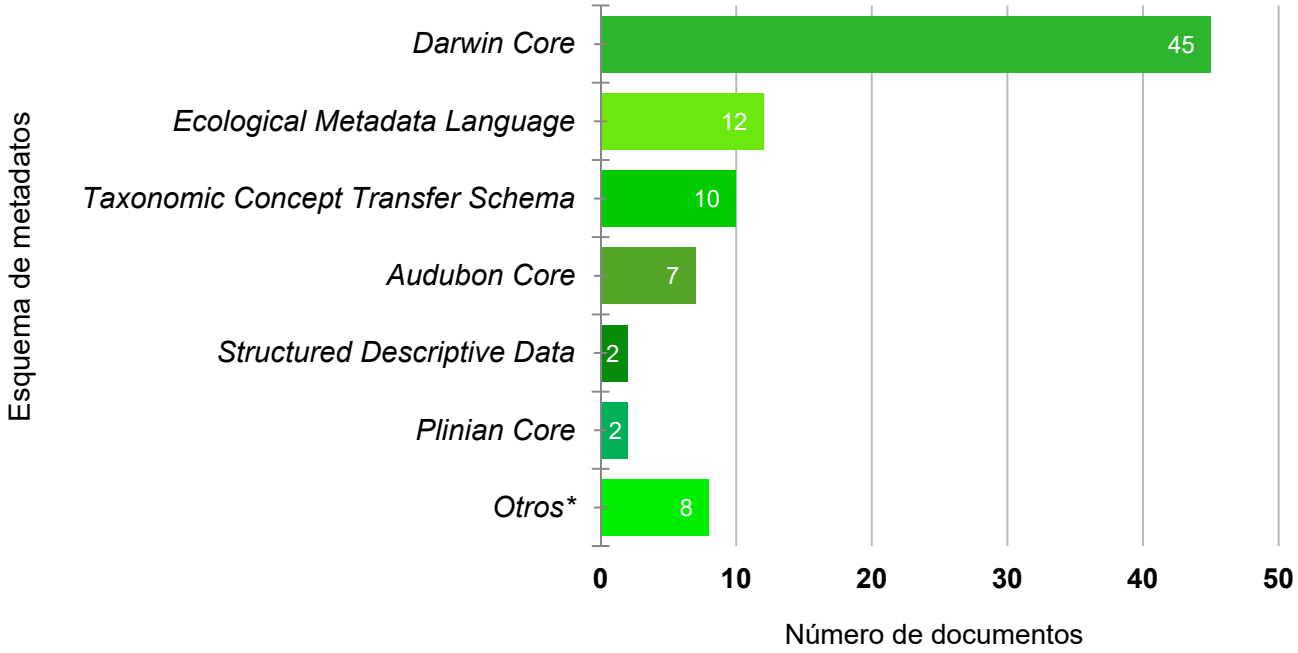


Figura 20. Tipos de recursos web creados para el manejo de datos digitales de especies.

Por otro lado, 85 recursos digitales citados en los documentos de investigación son de uso general, es decir, no fueron diseñados específicamente para datos de especies, pero pueden ser utilizados en el manejo de información de biodiversidad. A continuación, se describen detalladamente las tendencias en los metadatos, estándares, datos ligados, sistemas de organización del conocimiento e identificadores presentes en las investigaciones analizadas.

3. 1 Metadatos

Los textos científicos analizados en este trabajo, hacen referencia a 17 esquemas de metadatos, donde se incluyen aquellos diseñados para describir y estandarizar contenido digital de Internet o datos geográficos. El 82% (n=14) de estos recursos web, son específicos para el manejo de datos de especies (figura 21).



Otros* conjunto de esquemas de metadatos citados por una sola investigación.

Figura 21. Esquemas de metadatos más significativos para el manejo de información digital de especies.

Los esquemas de metadatos más relevantes, respecto al número de citas en los documentos analizados fueron: *Darwin Core*, *Ecological Metadata Language*, *Taxonomic Concept Transfer Schema*, *Audubun Core*, *Structured Descriptive Data* y *Plinian Core.* Ocho esquemas de metadatos son citados sólo por un documento de investigación, entre los que se encuentran: *Ecological Modeling Language* (Triebel, Hagedorn y Rambold, 2012), *EPGRIS3 Trait Data Standard* (Wieczorek *et al*., 2012), *TaxonX* (Agosti, y Egloff, 2009), *Herbarium Information Standards and Protocols for Interchange of Data* (Holetschek *et al*., 2012), *GBIF Metadata Profile* (Chavan y Penev, 2011), *Species Profile Model* (Parr *et al*., 2014), *WDCM minimum datasets* y *WDCM recommended datasets* (Wu *et al*., 2016).

3. 2 Estándares o especificaciones, lenguajes y protocolos

Se identificaron 51 estándares, de los cuales siete (13.72%) corresponden a recursos digitales creados especialmente para el manejo de datos de especies y 44 (86.27%) a especificaciones de uso general. Los estándares más relevantes diseñados en biodiversidad pertenecen principalmente a protocolos para transferir información entre colecciones digitales distribuidas en la *World Wide Web* (tabla 11).

Tabla 11

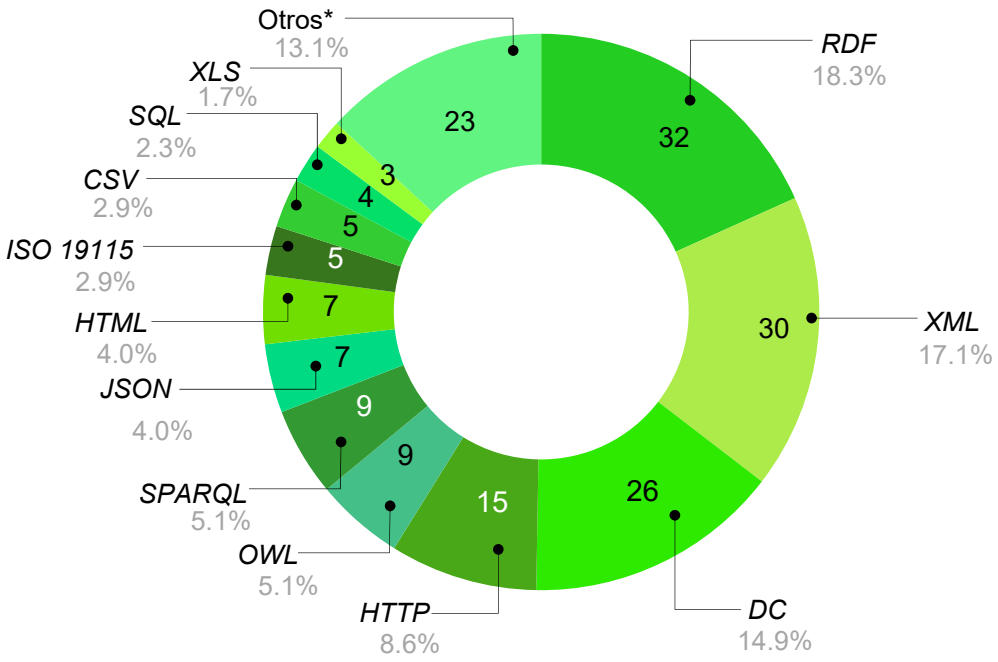
Estándares más relevantes para el manejo de datos digitales de especies

Estándar	Descripción	Citas
TDWG Access Protocol for Information Retrieval	Protocolo web para realizar consultas en bases de datos distribuidas con estructura física y lógica variable.	Agosti, y Egloff, 2009; Holetschek *et al*., 2009; Goddard *et al*., 2011; Bach *et al*., 2012; Holetschek *et al*., 2012; Candela *et al.*, 2015 y Kelbert *et al*., 2015
Darwin Core Archive	Formato para publicar conjuntos de datos y metadatos de taxonomía, presencia de especies o eventos de muestreo en colecciones digitales de biodiversidad. Es un estándar de acuerdo a los lineamientos del esquema de metadatos *Darwin Core.*	Goddard *et al*., 2011; Triebel, Hagedorn y Rambold, 2012; Tzitzikas *et al*., 2013; Costello y Wieczorek, 2014; Laurenne *et al*., 2014; Parr *et al*., 2014; Kelbert *et al*., 2015; Parr *et al*., 2016; Koureas *et al*.; 2016 y Rees *et al*., 2017
Distributed Generic Information Retrieval	Protocolo que vincula bases de datos independientes de la comunidad en una única colección virtual con capacidad de búsqueda basada en *XML* a una comunidad de fuentes de datos diferentes.	Jones, 2006; Holetschek *et al*., 2009; Goddard *et al*., 2011; Bach *et al.*, 2012; Candela *et al.*, 2015 y Kelbert *et al*., 2015
Biological Collection Access Service	Protocolo para transferir archivos, desarrollado para acceder a datos históricos de colecciones biológicas geográficamente distribuidas.	Holetschek *et al*., 2009; Goddard *et al*., 2011; Bach *et al*., 2012; Holetschek *et al*., 2012; Tschöpe *et al*., 2013 y Kelbert *et al*., 2015,
NeXML	Define la sintaxis para unidades taxonómicas operativas, matrices de estado de carácter y árboles y redes filogenéticas.	Vos *et al*., 2014 y Hugo *et al*., 2016

Fuente: Elaboración propia.

Los protocolos para el intercambio de información, permiten transferir y recuperar datos dentro de las colecciones digitales, por lo que son una importante contribución para el acceso y la distribución de información de biodiversidad en Internet (Costello y Vanden Berghe, 2006; Johnson, 2007; Holetschek *et al*., 2009). Sin embargo, la complejidad de los datos de especies, pone en evidencia la falta de estándares para codificar y representar algunos dominios de datos.

Los estándares de uso general citados en las investigaciones pertenecen a lenguajes para describir y consultar información, protocolos de intercambio de datos, formatos y normas internacionales. En total se identificaron 44 estándares que se describen en la figura 22.



Otros* hace referencia al conjunto de 23 estándares citados sólo por un documento de investigación.

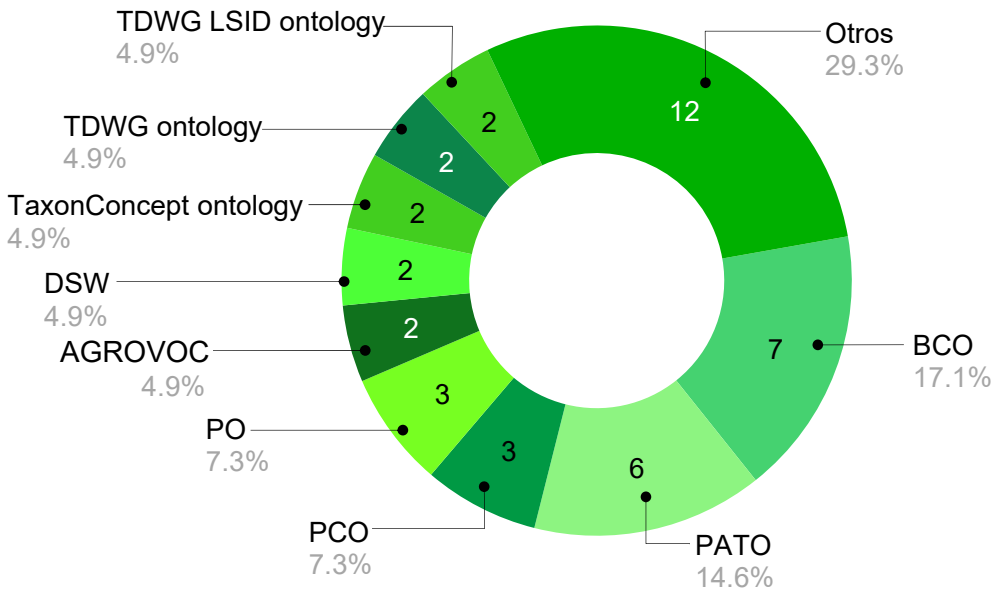
Figura 22. Estándares más relevantes para el manejo de información general.

Las principales especificaciones fueron: *Resource Description Framework* (RDF), *Extensible Markup Language* (XML) y *Dublin Core* (DC), (figura 22). La importancia de RDF radica en que es una de las principales herramientas para el intercambio de datos y posee un importante potencial para compartir datos de especies estructurados y semiestructurados en la web semántica (W3C, 2014). El formato de texto *Extensible Markup Language* (XML), es por su parte uno de los estándares más importantes en la actualidad para el intercambio electrónico de datos (W3C, 2016). Finalmente, *Dublin Core,* fue una especificación ampliamente citada para describir contenido digital, y puede ser empleada como un modelo de metadatos para información de especies, como el caso del esquema *Darwin Core* (DCMI, 2018; Wiezoreck *et al*., 2012). Dentro de este grupo de estándares, destacan la norma *ISO 19115* para describir conjuntos de datos geográficos, asociados a la ubicación geográfica de las especies (Cotter y Bauldock, 2000); y *Web Ontology Language* (OWL), que permite representar el conocimiento y sus relaciones a través de ontologías.

En otro orden, ocho investigaciones mencionan el uso de los códigos de nomenclatura biológica para el manejo de datos de especies. Estos artículos hacen referencia a los códigos de nomenclatura en general (Amanqui *et al*., 2016; Bingham *et al*., 2017), el *International Code of Nomenclature for algae*, *fungi and plants* (ICN), (Laurenne *et al*., 2014; Güntsch *et al*., 2017; Horton *et al*., 2017), el *International Code of Zoological Nomenclature* (ICZN), (Agosti, y Egloff, 2009; Laurenne *et al*., 2014; Chawuthai *et al*., 2016; Güntsch *et al*., 2017; Horton *et al*., 2017) y el *International Code of Botanical Nomenclature* (ICBN), (Agosti, y Egloff, 2009). Los códigos de nomenclatura han sido utilizados históricamente como un estándar para el manejo de información biológica, y deben de ser un elemento indispensable para las bases de datos de especies a nivel mundial, a pesar de ello, no son un elemento ampliamente citado en las investigaciones en el tema. Autores como Bingham *et al*. (2017) han identificado el uso del *Taxonomic Backbone* (TB) para normalizar datos taxonómicos dentro de colecciones digitales particulares, lo que ocasiona inconsistencias en la información de distintas colecciones y limita significativamente su interoperabilidad. Desde esta perspectiva, es importante considerar que las colecciones digitales de especies se indexan con el uso de nombres y éstos representan un enlace entre otros niveles de información en biodiversidad (Jayasiri *et al*., 2015).

3. 3 Sistemas de organización del conocimiento (SOC)

Los artículos analizados hacen referencia a 38 sistemas de organización del conocimiento, de los cuales 17 (46%) son de dominio general y 21 (54%) son propios de los datos digitales de organismos. Los SOC más relevantes para el manejo de información de especies se representan en la figura 23.



Otros* 12 sistemas de organización del conocimiento citados en una investigación.

Figura 23. Sistemas de organización del conocimiento para el manejo de datos de especies.

Dentro de los SOC enfocados a especies, la ontología más importante fue *Biological Collections Ontology* (BCO) diseñada para el manejo de información sobre colecciones biológicas, muestras ambientales, metagenómica y estudios ecológicos (BCO, 2018). El siguiente recurso digital más relevante fue *Phenotypic Quality Ontology*, un estándar enfocado en la descripción de rasgos, que es a su vez interoperable con ontologías del mismo dominio de información (PATO, 2018). Otros SOC destacados fueron las ontologías para describir rasgos de plantas *Plant Ontology* y *Flora Phenotype Ontology,* y recursos relacionados con la agricultura como *AGROVOC* y *NAL Agricultural Thesaurus*. Estos resultados reflejan una posible duplicación de esfuerzos en el desarrollo de sistemas de organización del conocimiento, debido a su similitud y dominio específico de información. Un conjunto de 12 sistemas de organización del conocimiento fue citado únicamente por un documento de investigación (otros*, ver figura 23), dentro de estos recursos web se encuentran vocabularios creados para el manejo de datos de colecciones digitales específicas como *Encyclopedia of Life Data Glossary* y *TraitBank Data Glossary* (Parr *et al.*, 2016).

Los documentos de investigación mencionan 19 sistemas de organización del conocimiento, no específicos para la información de especies. En este grupo, destaca la ontología *The Environment Ontology* (ENVO) para la descripción de datos ambientales y *Extensible Ontology for Observations* (OBOE), que describe semánticamente las observaciones

científicas. Los SOC no diseñados originalmente para los datos de especies, tienen la ventaja de normalizar información estrechamente relacionada con su contexto.

De acuerdo con el tipo de sistema de organización del conocimiento el 85.4% (n=41) del total corresponde a ontologías, mientras que el 7.3% (n=6) a glosarios y tesauros (figura 24A); esta misma tendencia se observa en el subconjunto formado por los sistemas de organización del conocimiento del dominio de biodiversidad, con 18 ontologías que representan el 81.8% del total (figura 24B).

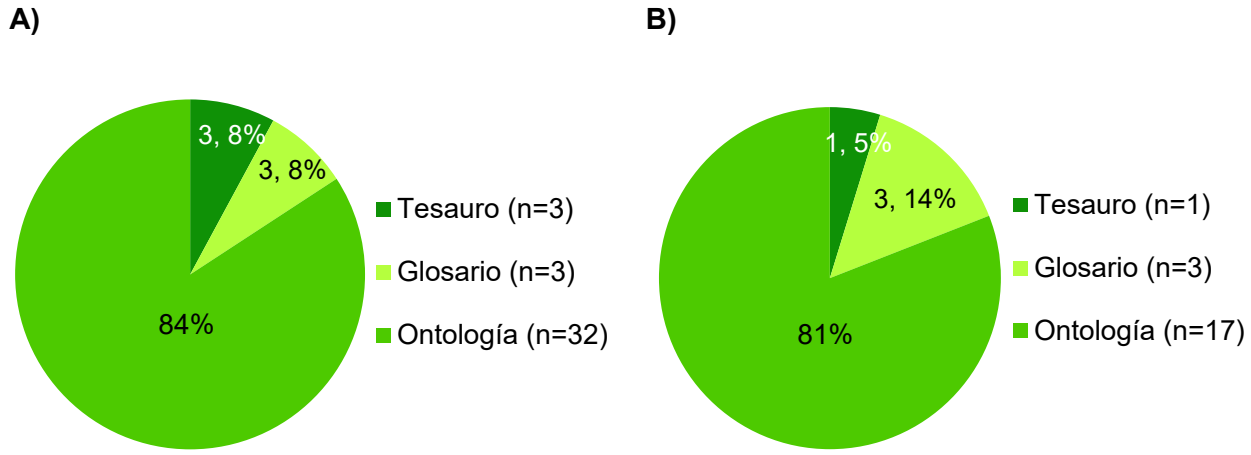


Figura 24. Tipos de sistemas de organización del conocimiento (SOC) A) Total de SOC (n=36), B) SOC de biodiversidad a nivel de especies.

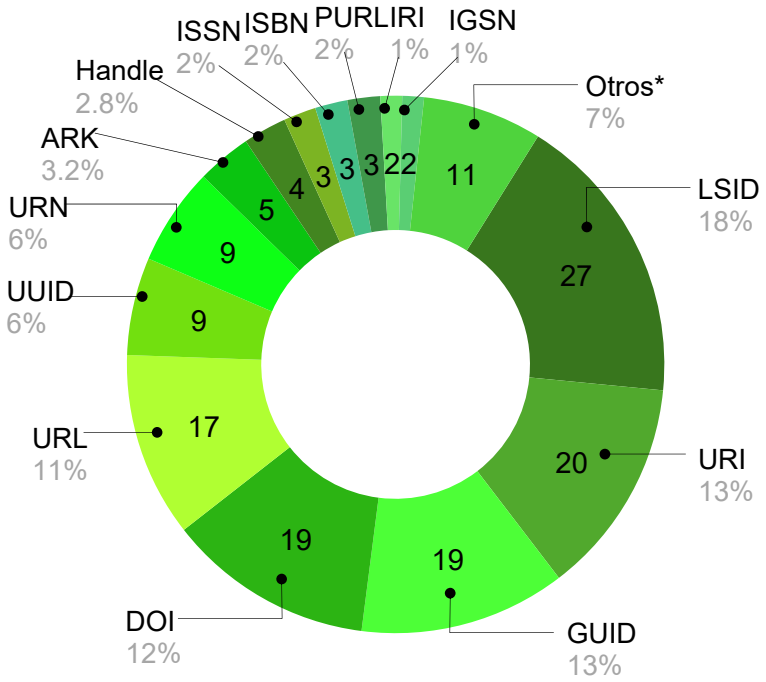
3. 4 Identificadores

Dentro de los artículos más relevantes se mencionan 25 identificadores, de los cuales cinco (20%) pertenecen a recursos digitales diseñados específicamente para el manejo de información de organismos. Los identificadores de biodiversidad fueron: *GBIF 'Triple IDs'* (Güntsch *et al*., 2017)*, AnnoSys 'tripleId'* (Suhrbier *et al*., 2017), *AphiaID* (Costello *et al*., 2013), *ION ID* (Rees *et al*., 2017) y *Darwin Core Triplet* (Guralnick *et al*., 2014). La mayoría de estos identificadores, a excepción de *Darwin Core Triplet,* fueron diseñados para repositorios específicos, es decir, sólo se utilizan para identificar registros en bases de datos individuales. El uso de identificadores restringidos a colecciones digitales particulares, representa un problema para el intercambio de datos, ya que puede conducir a la duplicación de los registros disponibles en la *World Wide Web*. De acuerdo con Costello y Vanden Berghe (2006) esta es una de las principales limitaciones en la información de biodiversidad, que puede resolverse mediante un consenso entre las

instituciones encargadas del manejo de datos, para utilizar identificadores persistentes de carácter global.

Los nombres taxonómicos han sido propuestos como identificadores para el manejo de información digital de especies (Costello, Horton y Kroh, 2018). Sin embargo, se descarta su uso debido a sus constantes actualizaciones (Huang y Qiao, 2011). Los nombres científicos han sido utilizados históricamente en la catalogación de especies y son un elemento común en toda la informática de la biodiversidad (Costello y Vanden Berghe, 2006; Sarkar, 2007), pero aún son inconsistentes entre las bases de datos actuales y se necesitan más investigaciones para poder explotar su potencial como un sistema de indexación de datos biológicos disponibles en línea (Thessen y Patterson, 2011; Chawuthai *et al*., 2016).

En lo que se refiere a los identificadores de uso general, se registraron 20 recursos web de este tipo, tal como se ilustra en la figura 25. Los identificadores más destacados de este conjunto son: el *Uniform Resource Identifier* (URI), el *Globally Unique Identifier* (GSID) y el *Digital Object Identifier* (DOI). Dentro de estos recursos, se situó al *Life Sciences Identifier* (LSID), que a pesar de ser una herramienta para el manejo de datos biológicos, no fue diseñado específicamente para información de especies.



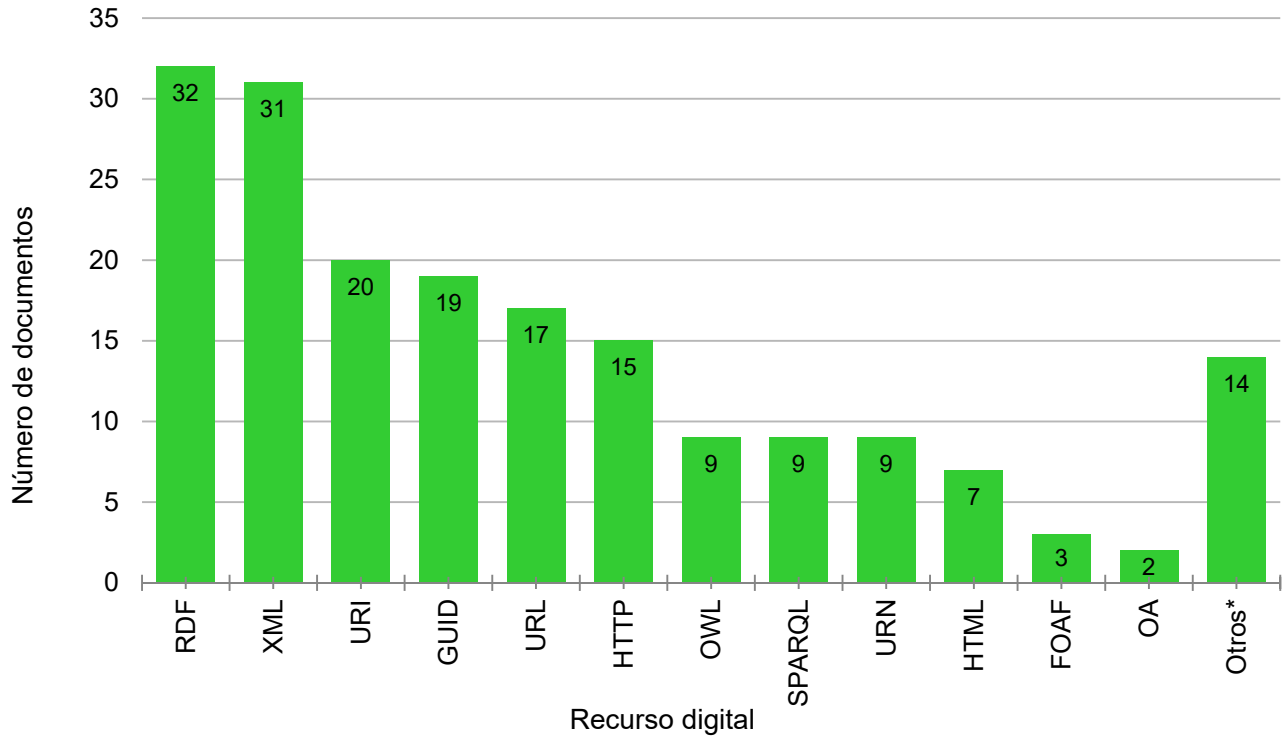
Otros* Conjunto de identificadores mencionados sólo por una investigación.

Figura 25. Identificadores para el manejo de información digital no específicos de especies.

Los principales identificadores observados en la figura 25 son conocidos como identificadores persistentes, ya que permiten identificar inequívocamente un objeto digital disponible en línea, a pesar de cambios en su ubicación, además tienen la particularidad de ser únicos y de carácter global y facilitan la vinculación de información de especies como datos ligados en la *World Wide* Web, con las ventajas que esto conlleva para el manejo de la información (Laurenne *et al*., 2014).

3. 5 Datos ligados

Los datos ligados son una forma de publicación de información en la *World Wide Web*, que implica el uso de un conjunto de recursos web de las categorías descritas anteriormente. De los 131 recursos digitales identificados en esta tesis, el 38.93% (n=51) se encuentra relacionado con la publicación de datos vinculados en la web semántica (figura 26). De acuerdo a su tipo, tres corresponden a esquemas de metadatos, seis a estándares, seis a identificadores y 35 a sistemas de organización del conocimiento. El 39.21% (n=20) fue diseñado para el manejo de datos de especies. La figura 26 representa las tecnologías más relevantes identificadas en los artículos de investigación.



Otros* grupo de tecnologías citadas sólo por un artículo de investigación.

Figura 26. Recursos web relativos a la publicación de datos ligados de especies.

Los recursos más significativos para la gestión de datos en la web 3.0 son: la especificación *RDF*, el estándar XML, los identificadores URI, GUID y URL, el estándar *Hypertext Transfer Protocol* (HTTP) y el lenguaje OWL (figura 26). Cabe señalar que estos recursos son de uso general y algunos han sido diseñados específicamente para la web semántica, como es el caso de las tecnologías RDF, SPARQL y OWL (W3C, 2015a). Este resultado, denota la importancia de los datos ligados en las investigaciones sobre el manejo de información de organismos, y su implementación tiene la ventaja de habilitar la automatización de procesos con el uso de la inteligencia artificial (Parr y Thessen, 2018). En otro orden, los recursos web diseñados para modelar datos de especies dentro de la web 3.0 pertenecen en su mayoría a sistemas de organización del conocimiento de tipo ontología, que han sido descritos en la sección 3.3.

3. 6 Datos abiertos

La publicación de datos abiertos vinculados dentro de la web semántica, necesita del libre acceso a la información disponible en la *World Wide Web*. Los datos abiertos se encuentran en el 69.5% (n=66) de las investigaciones en el manejo de datos de biodiversidad a nivel de especies. La figura 27, refleja el número de artículos publicados por año en el tema.

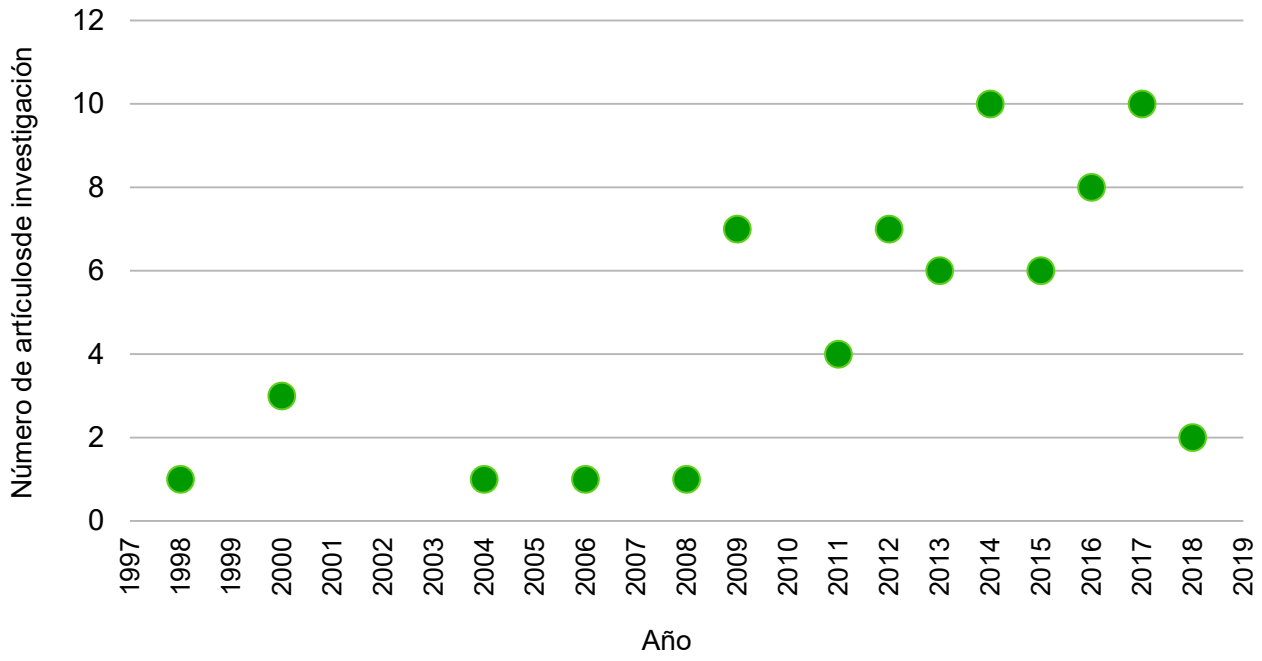


Figura 27. Año y número de artículos que mencionan el acceso abierto a los datos de especies disponibles en Internet.

Los documentos científicos, muestran que el acceso abierto es reconocido desde las primeras investigaciones. Tal como se aprecia en la figura 26, el tema ha cobrado notoriedad en los últimos años, a pesar de ser una práctica propuesta desde la conceptualización de las primeras bases de datos de especies en línea. El aumento en el número de investigaciones, puede responder a la necesidad actual de publicar los datos que fundamentan a las investigaciones y el movimiento de la ciencia abierta en general (Smith *et al*., 2009). Por otra parte, se identificó que un amplio número de autores recomiendan el uso de las licencias *Creative Commons (CC)* para publicar datos de diversidad biológica (Hugo *et al*., 2016; Costello *et al*., 2013; Groom *et al*., 2016; Smith *et al*., 2009; Chavan y Penev, 2011; Triebel, Hagedorn y Rambold, 2012). Algunos autores reconocen que el acceso abierto en biodiversidad puede resultar contraproducente, al tratarse de organismos en peligro de extinción, de interés comercial o considerados casos especiales (Cotter y Bauldock, 2000; Moritz *et al*., 2011). No obstante, de acuerdo con Moritz *et al*. (2011) los datos sensibles de especies representan una mínima proporción y su retención debe ser debidamente justificada.

La disponibilidad de los datos de especies se relaciona directamente con los datos oscuros, es decir, aquellos que no se encuentran digitalizados, o bien, no son publicados en colecciones de la *World Wide Web* (Edwards, Lane y Nielsen, 2000; Moritz *et al.,* 2011; Turnhout y Boonman-Berson, 2011). Especialmente, los datos producto de las investigaciones contemporáneas no se encuentran disponibles, debido a que son retenidos por sus creadores, por falta de conocimiento, capacidades técnicas, arraigo y por falta de atribución o reconocimiento a su trabajo (Moritz *et al.,* 2011). Se ha propuesto, que la distribución de los datos de especies debe ser reconocida formalmente como una publicación revisada por pares, citada e incluida en las métricas que evalúan la producción de los científicos (Costello, 2009; Costello *et al*., 2014). Además, existen pronunciamientos sobre la liberación obligatoria de los datos empleados en las investigaciones, tras ser publicadas en revistas (Bowker, 2000; Chavan y Penev, 2011). Sin embargo, aún se debe establecer una forma de publicación para los datos de especies, que elimine las limitaciones sociotécnicas, socioculturales, técnico-infraestructurales, políticas y legales, relacionadas directamente con el acceso abierto en biodiversidad (Agosti, y Egloff, 2009; Moritz *et al.,* 2011).

No se encontró una clara relación entre las investigaciones sobre los datos abiertos de especies y aquellas que tratan el uso de los datos ligados, ya que el 51.5 % (n=34) de los documentos que describen la publicación de información de especies en acceso abierto, no hacen referencia a los datos abiertos vinculados o a la web 3.0.

4. Implementación de recursos web en colecciones digitales de especies

La integración de la información de especies presente en la *World Wide Web* depende de la correcta aplicación de los recursos digitales (Walls *et al*., 2014). En esta sección, se representa la interoperabilidad básica entre los repositorios que reúnen datos de todos los grupos conocidos de seres vivos, con base en el uso de recursos electrónicos compartidos.

4. 1 Esquemas de metadatos

La interoperabilidad básica de las colecciones digitales de especies con base en los esquemas de metadatos resulta en 12 de 17 bases de datos que utilizan recursos web comunes, lo que representa el 70.58% del total. La integración e intercambio de datos entre las colecciones que conforman este conjunto se ilustran en la figura 28.

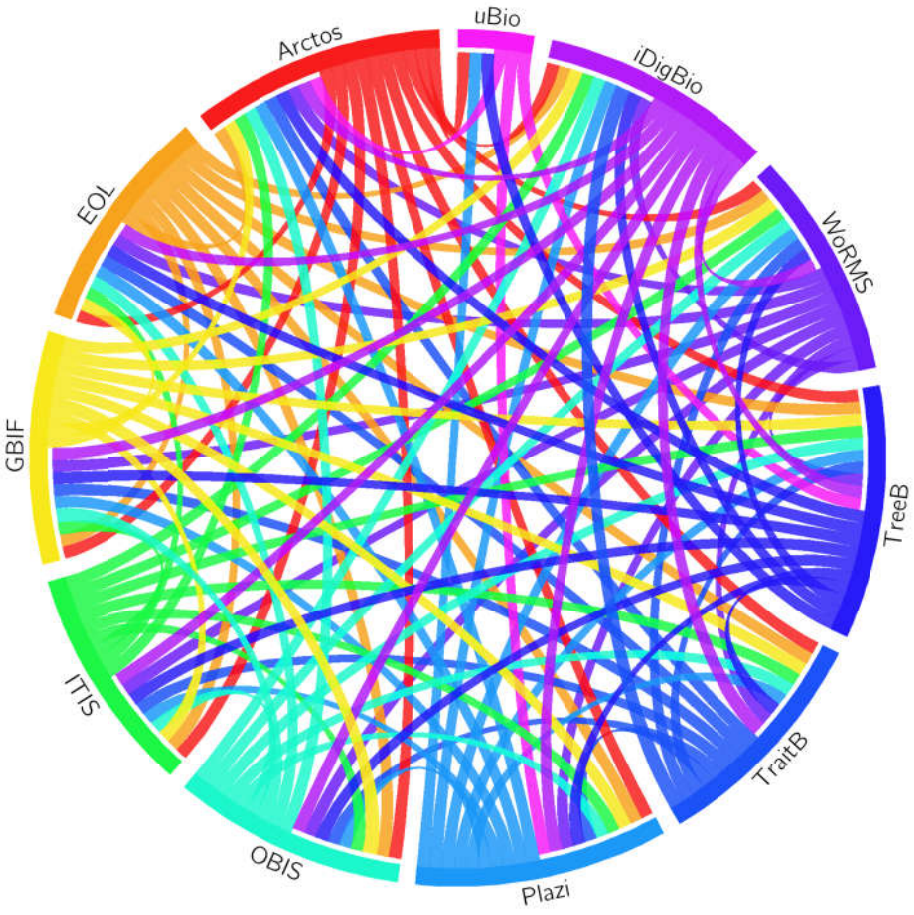


Figura 28. Interoperabilidad básica de las colecciones digitales de especies que emplean esquemas de metadatos comunes.

Dentro de este panorama, se observa que la mayoría de repositorios se encuentran conectados de forma equivalente. El menor número de conexiones se encontró en la base de datos *uBio* (The Marine Biological Laboratory, 2018), mientras que, la más comunicada fue *Arctos* (2018). En este conjunto se descartaron las colecciones *Global Invasive Species Database* (GSID), (IUCN, 2018b), *Global Register of Introduced and Invasive Species* (GRIIS), (GIASIPartnership, 2018), *NCBI Taxonomy* (NCBI, 2018b) y *Open Tree of Life* (OTL, 2018) que no especifican el uso de algún tipo de esquema de metadatos para describir su información.

En total, 11 de las 12 colecciones digitales se comunican con el uso de *Darwin Core*, un recurso diseñado para facilitar la recuperación e integración de los datos primarios que documentan la presencia de taxones (Johnson, 2007; Wieczorek *et al*., 2012). El reconocimiento y la adopción de este esquema, reflejan que es un caso de éxito en informática de la biodiversidad, ya que se encuentra en las investigaciones más relevantes y logra unificar e integrar datos de distintas colecciones disponibles en línea. A pesar de ello, la diversidad de información y los principales dominios de datos de especies, evidencian que este esquema no satisface todas las necesidades de la información en biodiversidad; tal como lo describen Groom *et al*. (2017) para el caso de las especies invasoras.

Los metadatos son fundamentales para que los usuarios de los repositorios comprendan el significado de los datos y cuenten con información suficiente para conocer las condiciones en que se tomaron. Permiten describir adecuadamente la información y comparar, explorar e integrar distintas fuentes distribuidas, por lo que son una herramienta que debería especificarse en el conjunto total de repositorios digitales de especies identificado en esta tesis (Cotter y Bauldock, 2000).

4. 2 Estándares

La aplicación práctica de los estándares dentro de las bases de datos de especies, resulta en 14 colecciones que poseen al menos una especificación en común para normalizar sus datos. La figura 29 muestra la capacidad de intercomunicación de este conjunto de repositorios.

.

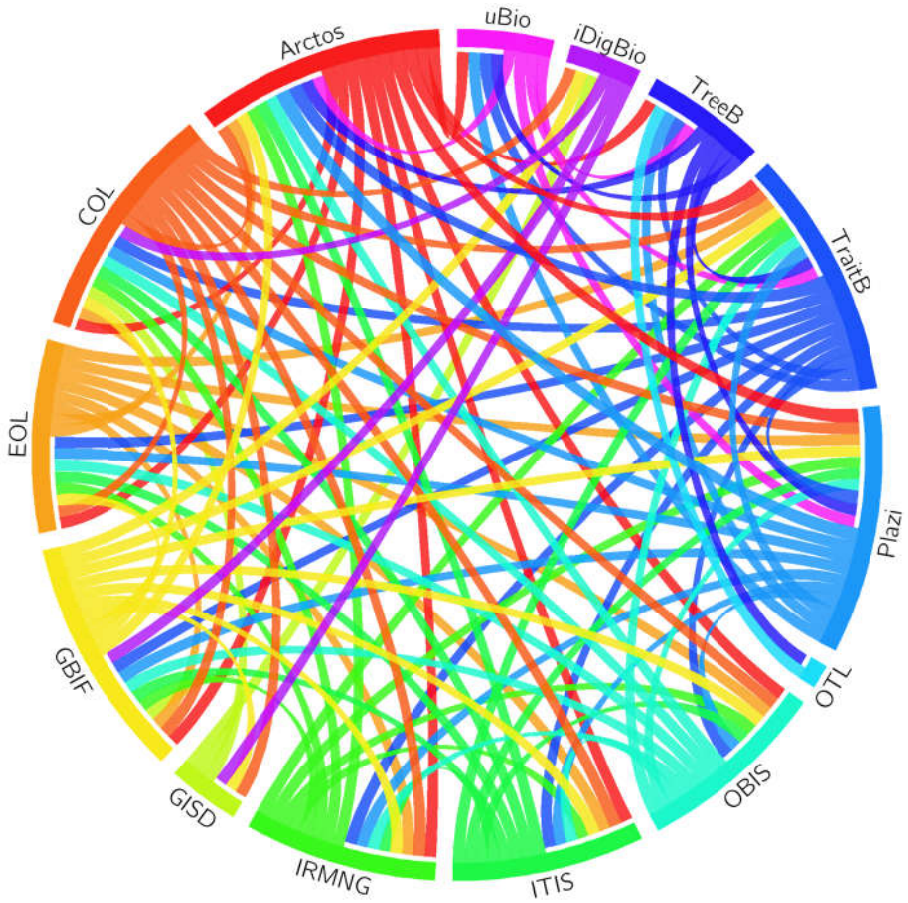


Figura 29. Interoperabilidad básica de las colecciones digitales de especies respecto al uso de estándares comunes.

Los repositorios que comparten el mayor número de estándares fueron *Plazi* (Agosti, s. f.), *Encyclopedia of Life* (EOL, 2018a), *TraitBank* (EOL, 2018b), *Catalogue of Life* (2018) y *Arctos* (2018), mientras que las colecciones *OTL* (NSF, 2018) *y Global Invasive Species Database* (IUCN, 2018b) presentaron el menor número de especificaciones comunes. En este análisis se omitieron las bases de datos *NCBI* (2018b) y *Global Register of Introduced and Invasive Species* (*GRIIS*), (GIASIPartnership, 2018), que no precisan el uso de estándares; y *World Register of Marine Species* (Flanders Marine Institute, 2018) que no comparte normas con otras colecciones del conjunto. El estándar más relevante fue *Darwin Core Archive*, que se encuentra en 8 de las 17 colecciones: GBIF (GBIF.org, 2018b), *Catalogue of Life* (COL, 2018), *Ocean Biogeographic Information Systems* (OBIS), (IODE, 2018), *EOL* (EOL, 2018a), *Plazi* (Agosti, s. f.), *The Interim Register of Marine and Nonmarine Genera* (Rees, 2018), *TraitBank* (EOL, 2018b) y *Arctos* (2018).

La norma *ISO 3166* para datos geográficos es empleada en *GBIF* (GBIF.org, 2018b), *COL* (2018), *Integrated Digitized Biocollections* (iDigBio, 2018) y *Global Invasive Species Database* (IUCN, 2018b). Las bases de datos *uBio* (The Marine Biological Laboratory, 2018), *TreeBASE* (2018), *Plazi* (Agosti, s. f.), *TraitBank* (EOL, 2018b) y *Arctos* (2018) utilizan la especificación *RDF*. Estas colecciones digitales se contextualizan semánticamente dentro de la web 3.0, lo que representa un paso importante para el manejo, análisis e interpretación de grandes cantidades de datos de especies. Las colecciones *GBIF* (GBIF.org, 2018b), *COL* (2018), *OBIS* (IODE, 2018), *WoRMS* (Flanders Marine Institute, 2018) y *OTL* (NSF, 2018) utilizan al menos un estándar que no se comparte con alguna otra colección analizada, lo que limita su interoperabilidad.

Los formatos de descarga de datos son un tipo de estándar de carácter técnico, que permite al usuario manejar información de distintas fuentes. En este caso 15 (88.23%) de las 17 bases de datos proporcionan los mismos tipos de archivos de datos (figura 30).

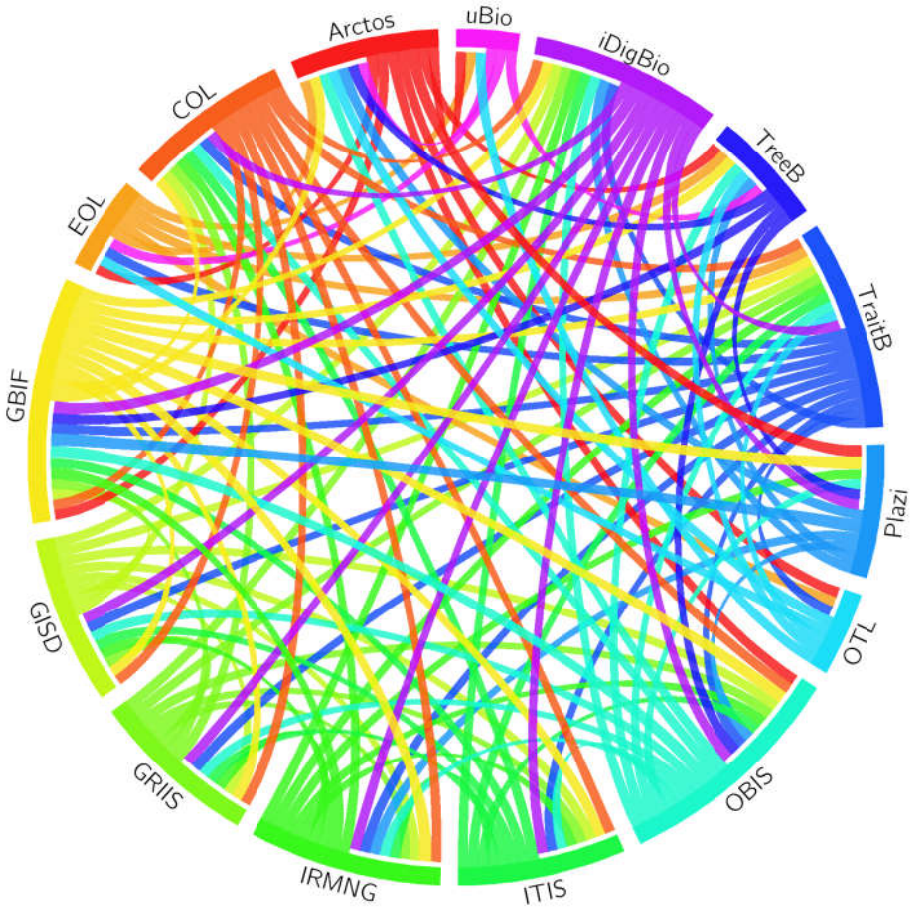


Figura 30. Interoperabilidad de las bases de datos de especies, respecto a los tipos archivos para la descarga de datos.

En este escenario, se descartaron únicamente dos repositorios que no permiten la descarga inmediata de los datos: *NCBI* (2018b) y *WoRMS* (Flanders Marine Institute, 2018). El uso de formatos comunes representa una ventaja para los curadores y usuarios de los datos, respecto al uso de *software* o programas necesarios para la gestión de la información. En general, las colecciones digitales presentan enlaces con al menos otras dos bases de datos. Aquellas que ofrecen la mayor variedad de formatos de descarga de datos son: *OBIS* (IODE, 2018), *GBIF* (GBIF.org, 2018b), *IRMNG* (Rees, 2018) e *iDigBio* (2018), y las colecciones con el menor número de formatos disponibles son *uBio* (The Marine Biological Laboratory, 2018) y *OTL* (2018) (figura 30).

El uso de estándares restringidos a colecciones representa un problema para el manejo de información de especies, ya que los datos deben de transformarse en información equivalente para poder ser intercambiados o integrados. De acuerdo con Martellos y Attorre (2012) el uso de estándares comunes debe ser una práctica de carácter obligatorio en los repositorios de biodiversidad.

4. 3 Sistemas de organización del conocimiento (SOC)

El análisis de la implementación de los sistemas de organización del conocimiento, refleja que no son un recurso web ampliamente utilizado en el subconjunto de colecciones que almacenan datos de todos los grupos de seres vivos. Se obtuvo que únicamente dos (11.17%) de las 17 colecciones digitales utilizan SOC comunes: *GBIF* (GBIF.org, 2018b) e *iDigBio* (2018). A su vez, colecciones como *ITIS* (2017), *TreeBASE* (2018), *The Interim Register of Marine and Nonmarine Genera* (IRMNG), (Rees, 2018) y *NCBI Taxonomy* (2018b), utilizan diccionarios y glosarios, que no se comparten con otros repositorios y tampoco representan las relaciones semánticas entre la información. A pesar de este resultado, Parr y Thessen (2018) describen el uso de ontologías en *Phenoscape* (2018), un claro ejemplo de una base de datos con razonamiento semántico que no fue incluida dentro del conjunto analizado. La falta de información descrita semánticamente en las colecciones digitales representa una limitante para el uso de la inteligencia artificial y la publicación de datos ligados de especies en la web semántica. Examinar el total de colecciones identificadas en esta tesis podría mostrar un panorama más alentador respecto al uso de los SOC en el manejo de datos de especies.

4. 4 Identificadores

La interoperabilidad en las colecciones de especies, con respecto al uso de identificadores muestra un panorama alentador, ya que 15 (88.23%) de las 17 bases de datos de seres vivos comparten al menos un recurso de este tipo para distinguir sus registros, como se muestra en la figura 31. En este conjunto se descartaron los repositorios *GRIIS* (GIASIPartnership, 2018) y *GISD* (IUCN, 2018b) que no precisan el uso de identificadores.

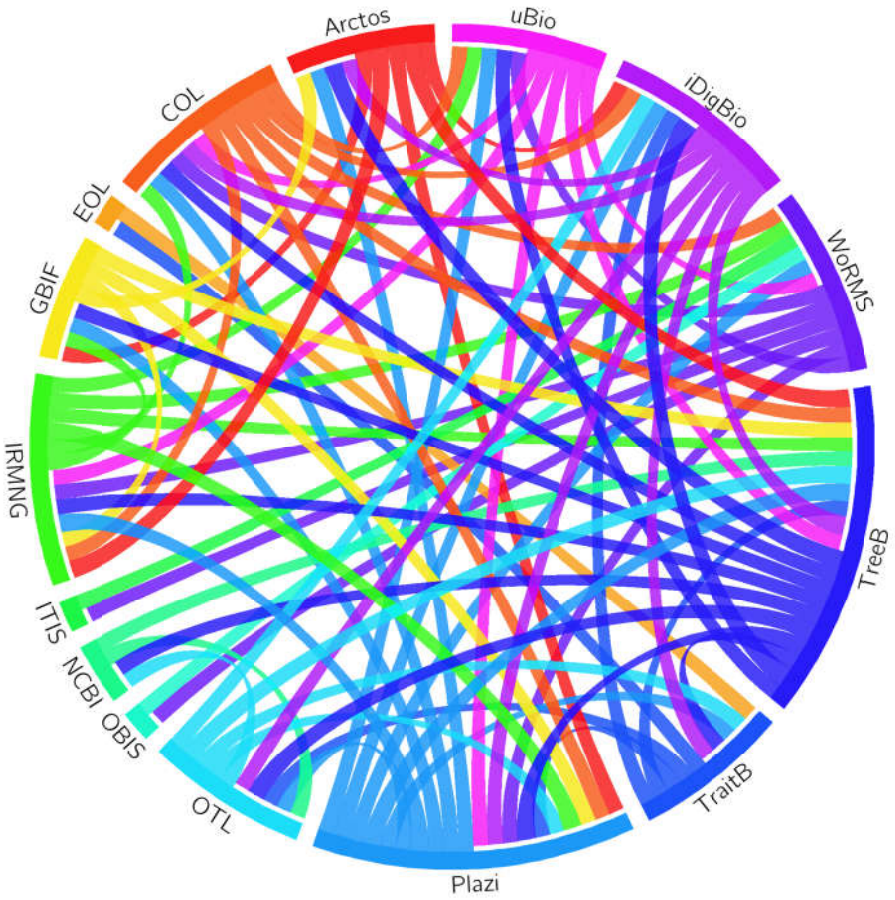


Figura 31. Interoperabilidad de las bases de datos de especies con respecto al uso de identificadores comunes.

Las bases de datos más comunicadas fueron: *Plazi* (Agosti, s. f.) y *TreeBASE* (2018), que utilizan una gran variedad de identificadores compartidos con al menos otras diez colecciones del conjunto. En contraste, las colecciones digitales menos relacionadas a otros elementos fueron *Encyclopedia of Life* (EOL, 2018a), *ITIS* (2017), *NCBI Taxonomy* (2018b) y *OBIS* (IODE, 2018). El análisis mostró que algunos identificadores fueron diseñados para colecciones específicas, pero han sido empleados en otras bases de datos que conforman el panorama (figura 31), como

es el caso del NCBI: txid (NCBI, 2018b) empleado por las colecciones *OTL* (2018) y *TreeBASE* (2018); y el *ITIS Taxonomic Serial Number* (ITIS TSN), (ITIS, 2017), utilizado por la colección *WoRMS* (2018). Los identificadores que representan el mayor número de enlaces son: *DOI*, *LSID* y el *Uniform Resource Identifier* (URI). El uso de cualquier tipo de identificadores compartidos representa un importante enlace para el descubrimiento de datos de organismos disponibles en la *World Wide Web* y reducen significativamente el riesgo de duplicar información.

4. 5 Datos ligados

La forma más elemental de los datos ligados o enlazados es el uso del protocolo HTTP, para crear enlaces entre recursos electrónicos identificados con URI que permiten descubrir información relacionada, con el uso de los denominados enlaces o *links.* La aplicación de los datos ligados en las colecciones digitales de especies se muestra en la figura 32

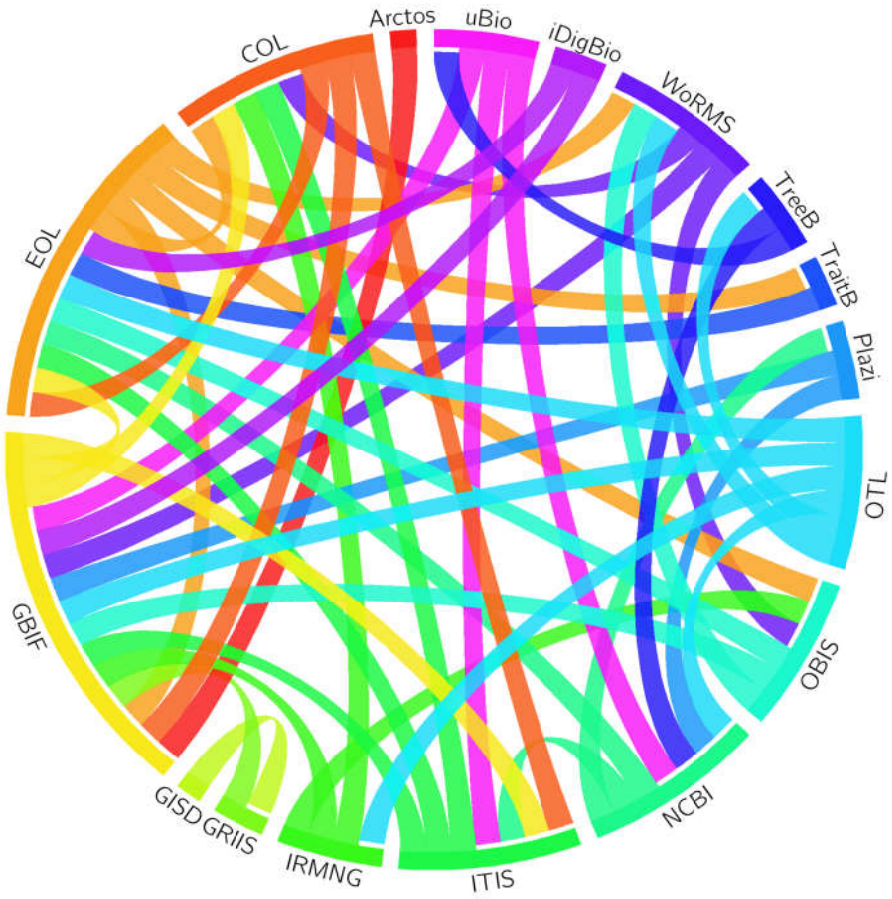


Figura 32. Datos ligados en colecciones digitales de especies con base en el uso de enlaces entre páginas web.

Dentro de este ámbito, podemos observar que las 17 bases de datos analizadas utilizan por lo menos un hiperenlace visible a otra colección digital. Los repositorios más relacionados son *EOL* (2018a) y *GBIF* (GBIF.org, 2018b), en ambos casos, el número de conexiones se explica por la cantidad de retroenlaces, es decir, otras bases de datos del conjunto son quienes se vinculan a estas colecciones. En total 12 de los 17 repositorios (70.58%) hacen referencia a *GBIF* (GBIF.org, 2018b) y ocho a *EOL* (2018a). En el caso contrario, los elementos menos conectados son *GRIIS* (GIASIPartnership, 2018), *GISD* (IUCN, 2018b), *Arctos* (2018), y *OTL* (NSF, 2018), esta última se relaciona a otras seis bases de datos, pero ninguna colección se liga a ella.

Uno de los principales hallazgos de esta investigación relacionado con la web 3.0, es el amplio desarrollo de ontologías en la comunidad de informática de la biodiversidad, las cuales permiten la publicación de datos abiertos ligados, sin embargo, falta adoptar por completo el total de recursos digitales relacionados. Por otro lado, el análisis de las bases de datos revela que en algunos casos ya se implementa la especificación RDF, lo que demuestra que algunas colecciones ya se encuentran en la web 3.0 y que el uso de la inteligencia semántica en la información de especies a nivel mundial es posible a mediano plazo.

4. 6 Datos abiertos

Las colecciones digitales de especies que almacenan información de todos los grupos de seres vivos se encuentran disponibles en acceso abierto en sus respectivos sitios web, lo que representa el 100% de las bases de datos evaluadas, e implica que cualquier usuario puede realizar una exploración previa de los datos e información presente en sus portales.

En lo que respecta al acceso y descarga de los registros, el 70.58% (n=12) de los repositorios permite la descarga inmediata de los datos. Las colecciones *GBIF* (GBIF.org, 2018b), *COL* (2018), *EOL* (EOL, 2018a), *TraitBank* (EOL, 2018b) y *Arctos* (2018) requieren de registro previo. En colecciones como *OBIS* (IODE, 2018) y *GBIF* (GBIF.org, 2018b) se promueve la atribución a los proveedores originales de los datos mediante el uso de citas.

El acceso abierto facilita la cooperación entre las bases de datos de biodiversidad y tiene importancia para la ciencia en general, pero también conlleva implicaciones en la actualización de los datos y en el ámbito legal respecto a los derechos de autor (Triebel, Hagedorn y Rambold, 2012). Otra de las ventajas del acceso abierto en las colecciones digitales de especies, es que propicia la creación de servicios y herramientas web de código abierto para el manejo de datos (Canhos *et al*., 2004).

5. Base de datos “Colecciones biodiversidad 2018”

Como resultado de esta tesis se creó la base de datos “Colecciones biodiversidad 2018”, que contiene información estructurada sobre el manejo de datos en colecciones digitales de especies. Los principales tipos de objetos de la base de datos fueron: registros bibliográficos, colecciones digitales y recursos electrónicos. Se conformó por 31 tablas de datos y los catálogos del SOC del Laboratorio de bioinformación que se describen en la tabla 12.

Tabla 12

Tablas del SOC del Laboratorio de bioinformación empleadas en esta tesis

Tabla	Descripción
Documento	Contiene la lista detallada de las secciones que contiene el documento de tesis.
Recuperación de información *Aalto*	Contiene la lista y descripción de las etapas del protocolo de recuperación de información de la Universidad de Aalto.
Digital Curation Centre	Contiene la lista y descripción de las etapas requeridas para la conservación y curación de los datos a lo largo de su ciclo de vida de acuerdo con el *Digital Curation Centre* (DCC).
CRAAPT	Contiene cada uno de los criterios para la selección de fuentes de información a partir de los criterios de la prueba "CRAAPT".
Idioma	Lista de idiomas basada en el código ISO 639-1 para la representación de nombres de lenguajes.
Región geográfica	Contiene una lista de las regiones geográficas a diferentes niveles incluyendo la lista de países del mundo, de acuerdo con la norma M49.
Tipo recurso	Contiene una lista de tipos de recursos electrónicos, para la clasificación de los objetos digitales recuperados.
Equivalencia	Contiene las equivalencias en los metadatos de las colecciones bibliográficas *WoS*, *Scopus*, *DOAJ* y *PubMed*.

El uso de los SOC permitió dar significado y contexto a la información, lo que facilitará su posterior consulta, recuperación y reúso a largo plazo por cualquier usuario. La base de datos fue diseñada de acuerdo al modelo relacional con el uso de registros y campos para almacenar entidades y sus atributos, respectivamente (figura 33).

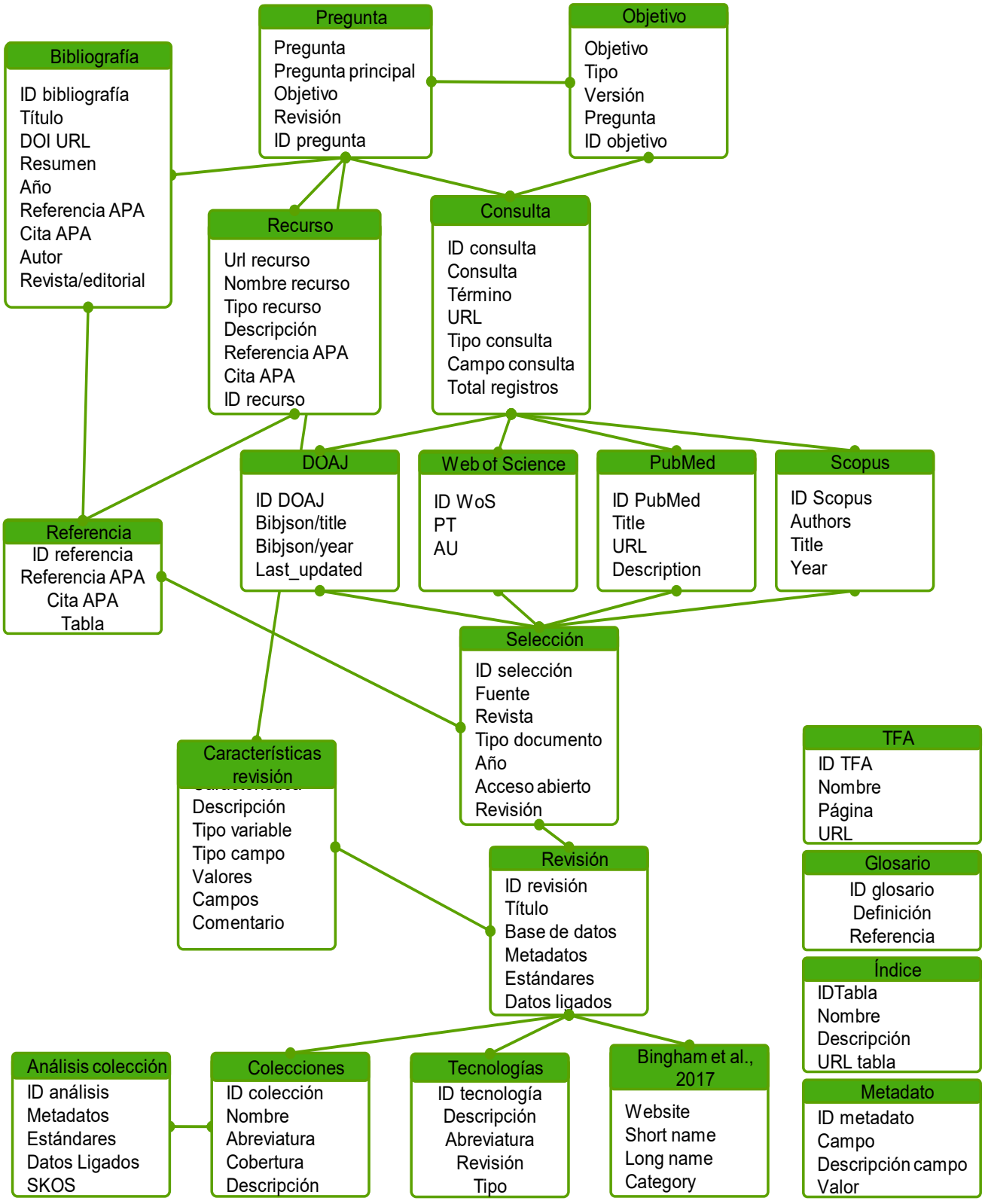


Figura 33. Representación de la estructura de la base de datos “Colecciones biodiversidad 2018”.

La base de datos se encuentra alojada en el sistema de información del Laboratorio de bioinformación de la Facultad de Ciencias (figura 34) y puede ser consultada por cualquier persona interesada.

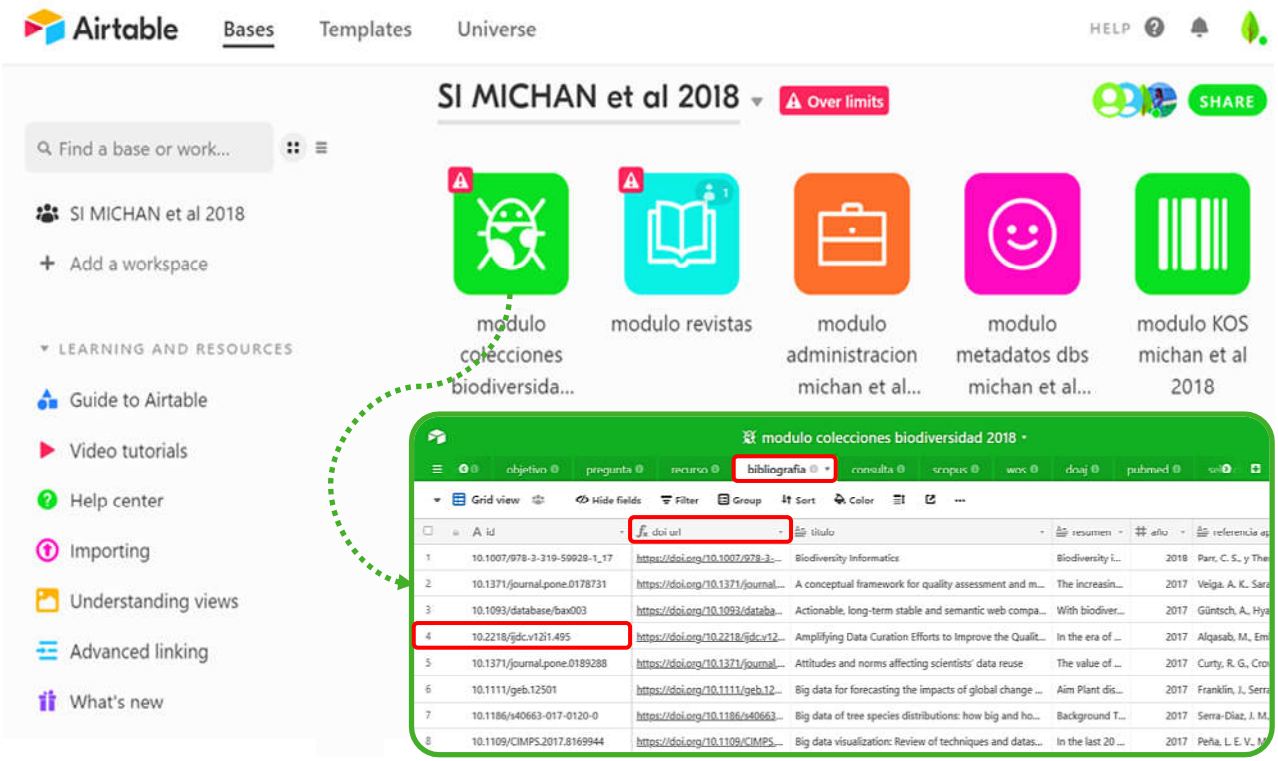


Figura 34. Vista general de la base de datos “Colecciones biodiversidad 2018”.

Conclusiones

1. Se identificaron 95 artículos de investigación sobre el manejo de información de especies, publicados en su mayoría en los últimos diez años, lo que refleja la importancia del tema en la práctica científica actual. El 96.84% de los artículos cita por lo menos a una colección digital de especies, mientras que los estándares, el acceso abierto y los metadatos son los temas mejor estudiados.
2. Esta tesis contribuye al conocimiento actual sobre la infraestructura de la informática de la biodiversidad, al reunir 72 colecciones digitales que preservan y publican datos de organismos en Internet. Este conjunto permitirá realizar estudios futuros que ayuden a resolver los problemas actuales de interoperabilidad; o bien puede ser empleado para conocer las fuentes de información disponibles para publicar o consultar datos mundiales de diversidad biológica.
3. Las 72 colecciones digitales de especies identificadas en esta tesis corresponden en su mayoría al dominio de información de taxonomía-nomenclatura y rasgos-datos descriptivos, se concentran en países desarrollados y no representan equitativamente a todos los grupos conocidos de organismos.
4. Las herramientas disponibles para representar información de especies, no cubren por completo las necesidades de todos los tipos de datos existentes, y corresponden en su mayoría a ontologías altamente especializadas y similares que pueden resultar redundantes.
5. Las bases de datos de seres vivos más relevantes, no se encuentran vinculadas equitativamente, ya que no comparten la mayoría de los recursos web empleados en el manejo de sus datos, y en algunos casos utilizan herramientas diseñadas para sus propias colecciones, lo que limita considerablemente su interoperabilidad.
6. El acceso abierto es uno de los temas más importantes para el manejo de datos en informática de la biodiversidad, con un amplio número de artículos de investigación disponibles libremente, un creciente número de investigaciones en el tema, importantes propuestas para la publicación de datos abiertos de especies en Internet y el reconocimiento hacia las colecciones digitales disponibles bajo los principios del acceso abierto.

7. Los datos ligados han cobrado notoriedad en el manejo de datos de especies en los últimos años, con un aumento en el número de artículos de investigación que recomiendan su uso y con el desarrollo de ontologías que permiten describir información. En la práctica, algunas colecciones ya implementan la especificación RDF y tecnologías relacionadas con la web semántica como HTTP y XML.
8. El estado del conocimiento respecto al uso y desarrollo de metadatos, estándares y datos ligados en el manejo de datos de diversidad biológica disponibles en colecciones digitales, evidencia la falta de bases de datos y recursos web para gestionar algunos dominios y tipos de información, la ausencia de consensos en la adopción de las herramientas existentes y el creciente interés en la publicación de datos abiertos y datos ligados de especies en la web semántica.

Perspectivas

- Los problemas de interoperabilidad entre las colecciones digitales de especies, deben de ser analizados de forma integral, involucrando a los autores, las colecciones digitales, los curadores, los desarrolladores de herramientas informáticas y los usuarios finales de la información.
- La implementación de metadatos, estándares y datos ligados, debe de examinarse en el conjunto más completo posible de bases de datos de especies, considerando su dominio de información.
- El manejo de datos digitales debe ser analizado a escala local y regional para identificar el progreso actual de la infraestructura disponible y mejorar así la gestión de datos de diversidad biológica a gran escala.

Glosario

Aplicación web (*Web application*) Sistemas distribuidos en Internet que brindan servicios a los usuarios a través de un servidor, su alcance y complejidad varía de pequeña a gran escala (UNESCO, 2018).

Complemento (*plugin*) Módulo que agrega una nueva característica o función a otro programa informático (Wikidata, 2019). Permite automatizar tareas como la búsqueda de artículos y referencias desde cualquier página web dentro de un navegador.

Consulta (*query*) Secuencia de comandos que definen una tarea de búsqueda dentro de una colección o base de datos (Ison *et al*., 2013).

Datos (*data*) Representación de observaciones, descripciones o mediciones sobre un objeto, fenómeno o evento que se registra generalmente de forma estandarizada y específica (Moritz *et al*., 2011).

Diccionario de datos (*Data dictionary*) Conjunto formal de términos usados para describir información almacenada en una colección o base de datos (DCC, 2018).

***Digital Object Identifier* (*DOI*)** Estándar internacional de carácter permanente que distingue una entidad presente en redes digitales; proporciona un enlace a su información actual, y puede resolverse en datos relacionados, como direcciones de correo, otros identificadores, metadatos descriptivos asociados, etc. (International DOI Foundation, 2017).

Dominio (*domain*) Disciplina científica, campo de estudio, o comunidad discursiva agrupada por nexos comunes que se representan de forma dinámica (Hjørland; Albrechtsen, 1995, en Tirador Ramos, 2010).

Enlace (*link*) Secuencia de caracteres que hacen referencia a otros recursos y representan información adicional en uno o varios servidores; también denominado hiperenlace (RAE, 2019; Wikidata, 2019).

Formato (*format*) Diseño establecido para representar y estructurar datos en un archivo de computadora, BLOB (*Binary Large Objects*), cadena, mensaje, etc. (Ison *et al*., 2013).

Nacido digital (*born-digital*) Archivos creados en formato digital, que no provienen de un soporte físico (DCC, 2018).

Objeto digital (*digital objects*) Conjunto de archivos (texto, imágenes o sonido), identificadores y metadatos relacionados que representan entidades físicas o abstractas. Pueden ser resultado de la combinación de objetos digitales para crear materiales complejos como los sitios web (International DOI Foundation, 2017; DCC, 2018).

RSS (*Really Simple Syndication*) Documento en formato XML para la redifusión web de los metadatos y contenidos de un sitio (UserLand Software, 2002).

Servicio web (*web service*) Recurso de tipo *software* cuyas funciones se encuentran disponibles en Internet (NCIT, 2019).

Programa de computadora (*Computer Program*) Conjunto de instrucciones codificadas para procesar datos, realizar operaciones o resolver problemas lógicos en un ordenador (NCIT, 2019).

***Uniform Resource Identifier* (URI)** Secuencia compacta de caracteres que identifica un recurso digital que forma parte de la *World Wide Web* (Berners-Lee *et al*., 1998).

***Uniform Resource Locator* (URL)** Subconjunto de URI que identifica un recurso digital a través de su ubicación en la red (Berners-Lee *et al*., 1998).

Referencias

Aalto University. (2018). Guide to information retrieval. Recuperado de http://libguides.aalto.fi/c.php?g=410678&p=2797969

Agosti, D. (s. f.). PLAZI. Recuperado de http://www.plazi.org

Agosti, D., y Egloff, W. (2009). Taxonomic information exchange and copyright: the Plazi approach. *BMC Research Notes*, 2(1), 53. doi:10.1186/1756-0500-2-53

Ahmad, A., Cuomo, S., Wu, W., y Jeon, G. (2019). Intelligent algorithms and standards for interoperability in Internet of Things. *Future Generation Computer Systems*, 92, 1187–1191. doi:10.1016/j.future.2018.11.015

Airtable. (2018). Base de datos Airtable. Recuperado de https://airtable.com/

Amanqui, F., Verborgh, R., Mannens, E., Van de Walle, R., y Moreira, D. (2016). Using Spatiotemporal Information to Integrate Heterogeneous Biodiversity Semantic Data. *Web Engineering*, 525–530. doi:10.1007/978-3-319-38791-8_41

ANDS. (2018). Metadata [Portal]. Recuperado de https://www.ands.org.au/guides/metadata-working

Anónimo. (2018). Highlight Tool (v. 60) [Complemento]. Recuperado de https://bit.ly/2KN7x3O

APA. (2018). APA Style. Recuperado de http://www.apastyle.org/learn/index.aspx

Arctos. (2018). Arctos Data Portal. Recuperado de https://arctosdb.org

Bach, K., Schäfer, D., Enke, N., Seeger, B., Gemeinholzer, B., y Bendix, J. (2012). A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. *Ecological Informatics*, 11, 16–24. doi:10.1016/j.ecoinf.2011.11.008

Baskauf, S. J., Wieczorek, J., Deck, J., y Webb, C. O. (2016). Lessons learned from adapting the Darwin Core vocabulary standard for use in RDF. *Semantic Web*, 7(6), 617–627. doi:10.3233/sw-150199

BCO. (2018). Biological Collections Ontology (BCO). Recuperado de http://www.obofoundry.org/ontology/bco.html

Beckstein, C., Böcker, S., Bogdan, M., Bruehlheide, H., M. Bücker, H., Denzler, J., … Zimmermann, W. (2014). Explorative Analysis of Heterogeneous, Unstructured, and Uncertain Data - A Computer Science Perspective on Biodiversity Research. *Proceedings of 3rd International Conference on Data Management Technologies and Applications*. doi:10.5220/0005098402510257

Berners-Lee, T., Fielding, R., Irvine, U. C., y Masinter, L. (1998). Uniform Resource Identifiers (URI): Generic Syntax [Estándar]. Recuperado de http://www.rfc-editor.org/rfc/rfc2396.txt

Berners-Lee, T. (2006). Linked data [Portal]. Recuperado de https://www.w3.org/DesignIssues/LinkedData.html

Bingham, H., Doudin, M., Weatherdon, L., Despot-Belmonte, K., Wetzel, F., Groom, Q., … Martin, C. (2017). The Biodiversity Informatics Landscape: Elements, Connections and Opportunities. *Research Ideas and Outcomes*, 3, e14059. doi:10.3897/rio.3.e14059

Bisby, F. A. (2000). The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*, 289(5488), 2309–2312. doi:10.1126/science.289.5488.2309

Bluhm, B., Watts, D., y Huettmann, F. (2010). Free Database Availability, Metadata and the Internet: An Example of Two High Latitude Components of the Census of Marine Life. *Spatial Complexity, Informatics, and Wildlife Conservation*, 233–243. doi:10.1007/978-4-431-87771-4_13

Bowker, G. C. (2000). Biodiversity Datadiversity. *Social Studies of Science*, 30(5), 643–683. doi:10.1177/030631200030005001

Brainerd, E. L., Blob, R. W., Hedrick, T. L., Creamer, A. T., y Müller, U. K. (2017). Data Management Rubric for Video Data in Organismal Biology. I*ntegrative and Comparative Biology*, 57(1), 33–47. doi:10.1093/icb/icx060

Bratková, E., y Kučerová, H. (2014). Knowledge organization systems and their typology. *Revue of Librarianship*, 25(2), 1-25. http://full.nkp.cz/nkkr/knihovna142_suppl/1402sup01.htm

Candela, L., Castelli, D., Coro, G., Lelii, L., Mangiacrapa, F., Marioli, V., y Pagano, P. (2015). An infrastructure-oriented approach for supporting biodiversity research. *Ecological Informatics*, 26, 162–172. doi:10.1016/j.ecoinf.2014.07.006

Canhos, V. P., Souza, S. D., Giovanni, R. D., y Canhos, D. A. L. (2004). Global Biodiversity Informatics: setting the scene for a “new world” of ecological forecasting. *Biodiversity Informatics*, 1(0). doi:10.17161/bi.v1i0.3

Castillo, M., Michán, L. y Martínez, A. (2014). La biocuración en biodiversidad: proceso, aciertos, errores, soluciones y perspectivas. *Acta botánica mexicana*, (108), 81-103. Recuperado en 04 de febrero de 2019, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-71512014000300006&lng=es&tlng=es

Catapano, T., Hobern, D., Lapp, H., Morris, R. A., Morrison, N., Noy, N., Schildhauer, M. y Thau, D. (2011). Recommendations for the Use of Knowledge Organisation Systems by GBIF. Biodiversity Information Facility, 49 pp. http://www.gbif.jp/v2/pdf/gbif_kos_whitepaper_v1.pdf

CDB. (1992). Convention on Biological Diversity [Documento]. Recuperado de https://www.cbd.int/doc/legal/cbd-en.pdf

Chavan, V., y Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, 12(Suppl 15), S2. doi:10.1186/1471-2105-12-s15-s2

Chavan, V. S., y Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*, 10(Suppl 14), S2. doi:10.1186/1471-2105-10-s14-s2

Chawuthai, R., Takeda, H., Wuwongse, V., y Jinbo, U. (2016). Presenting and preserving the change in taxonomic knowledge for linked data. *Semantic Web*, 7(6), 589–616. doi:10.3233/sw-150192

ChromeFans.org. (2017). Open SEO Stats (9.6.0.0) [Complemento]. Recuperado de chrome://extensions/?id=hbdkkfheckcdppiaiabobmennhijkknn

Clarivate Analytics. (2018a). Web of Science [v.5.29] [Colección]. Recuperado de https://clarivate.com/products/web-of-science/

Clarivate Analytics. (2018b). Colección principal de Web of Science Ayuda [Tutorial]. Recuperado de http://images.webofknowledge.com//WOKRS529AR7/help/es_LA/WOS/hp_search.html

COL. (2018). Catalogue of Life. [Colección]. Recuperado de http://www.catalogueoflife.org

CONABIO. (2008, 19 de diciembre). El Sistema Nacional de Información sobre Biodiversidad de México. [Portal]. Recuperado de: http://www.conabio.gob.mx/institucion/snib/doctos/acerca.html

CONABIO. (2015, 23 de febrero). Red Mundial de Información sobre Biodiversidad. Recuperado de: http://www.conabio.gob.mx/remib/doctos/remib_esp.html

Costello, M. J., y Vanden Berghe, E. (2006). “Ocean biodiversity informatics”: A new era in marine biology research and management. *Marine Ecology Progress Series*, 316, 203–214.

Costello, M. J. (2009). Motivating Online Publication of Data. *BioScience*, 59(5), 418–427. doi:10.1525/bio.2009.59.5.9

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q., y Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology y Evolution*, 28(8), 454–461. doi:10.1016/j.tree.2013.05.002

Costello, M. J., y Wieczorek, J. (2014). Best practice for biodiversity data management and publication. *Biological Conservation*, 173, 68–73. doi:10.1016/j.biocon.2013.10.018

Costello, M. J., Appeltans, W., Bailly, N., Berendsohn, W. G., de Jong, Y., Edwards, M., … Bisby, F. A. (2014). Strategies for the sustainability of online open-access biodiversity databases. *Biological Conservation*, 173, 155–165. doi:10.1016/j.biocon.2013.07.042

Costello, M. J., Horton, T., y Kroh, A. (2018). Sustainable Biodiversity Databasing: International, Collaborative, Dynamic, Centralised. *Trends in Ecology y Evolution*. doi:10.1016/j.tree.2018.08.006

Cotter, G. A., y Bauldock, B. T. (2000, September). Biodiversity Informatics Infrastructure: An Information Commons for the Biodiversity Community. *In VLDB* (pp. 701-704)

Daltio, J., y Medeiros, C. B. (2008). Aondê: An ontology Web service for interoperability across biodiversity applications. *Information Systems*, 33(7-8), 724–753. doi:10.1016/j.is.2008.02.001

Davies, T. G., Rahman, I. A., Lautenschlager, S., Cunningham, J. A., Asher, R. J., Barrett, P. M., … Donoghue, P. C. J. (2017). Open data and digital morphology. *Proceedings of the Royal Society B: Biological Sciences*, 284(1852), 20170194. doi:10.1098/rspb.2017.0194

DCC. (2018). Digital Curation Centre (DCC) [Portal]. Recuperado de http://www.dcc.ac.uk/

DCMI. (2018). Dublin Core Metadata Initiative (DCMI) [Sociedad]. Recuperado de http://dublincore.org/

De Pooter, D., Appeltans, W., Bailly, N., Bristol, S., Deneudt, K., Eliezer, M., … Hernandez, F. (2017). Toward a new data standard for combined marine biological and environmental datasets-expanding OBIS beyond species occurrences. *Biodiversity Data Journal*, 5, e10989. doi:10.3897/bdj.5.e10989

Digital Thoughts. (2018). Advanced URL Shortener (v. 22) [Complemento]. Recuperado de https://chrome.google.com/webstore/detail/advanced-url-shortener/jdcidaibeoijokhffgmihkojhjcbnbhd

DOAJ. (2018a). Directory of Open Access Journals [Colección]. Recuperado de https://doaj.org/

DOAJ. (2018b). DOAJ API [Aplicación]. Recuperado de https://doaj.org/api/v1/docs

Edwards, J. L. (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science*, 289(5488), 2312–2314. doi:10.1126/science.289.5488.2312

Edwards, J. L., Lane, M. A., y Nielsen, E. S. (2000). Interoperability of biodiversity databases: biodiversity information on every desktop. *Science*, 289(5488), 2312-2314.

Elasticsearch. (2018). Query String Query [Tutorial]. Recuperado de https://www.elastic.co/guide/en/elasticsearch/reference/1.4/query-dsl-query-string-query.html#query-string-syntax

Elsevier. (2018a). Scopus [Colección]. Recuperado de https://www.scopus.com/

Elsevier. (2018b). How do I search for a document? [Tutorial]. Recuperado de https://service.elsevier.com/app/answers/detail/a_id/11213/supporthub/scopus/

EOL. (2018a). Encyclopedia of Life (EOL) [Colección]. Recuperado de: http://eol.org

EOL. (2018b). TraitBank® [Colección]. Recuperado de: http://eol.org/info/516

Foster, E. C., y Godbole, S. (2016). Database systems: a pragmatic approach. *Apress*. doi:10.1007/978-1-4842-1191-5

FOSTER. (2018). Open Reproducible Research (FOSTER) [Portal]. Recuperado de https://www.fosteropenscience.eu/foster-taxonomy/open-reproducible-research

Franz, N. M., y Sterner, B. W. (2018). To increase trust, change the social design behind aggregated biodiversity data. *Database*, 2018. doi:10.1093/database/bax100

GBIF.org (2018a). Datos masivos de biodiversidad: GBIF.org sobrepasa los mil millones de registros de presencia de especies. *Global Biodiversity Information Facility*. Recuperado de: https://www.gbif.org/news/6jJnpi5mbC4kWM2yQ2KeYQ/datos-masivos-de-biodiversidad-gbiforg-sobrepasa-los-mil-millones-de-registros-de-presencia-de-especies

GBIF.org. (2018b). Global Biodiversity Information Facility [Colección]. Recuperado de https://www.gbif.org

GEO BON. (2019). Species Traits. Recuperado de: https://geobon.org/ebvs/working-groups/species-traits/

GIASIPartnership. (2018). Global Register of Introduced and Invasive Species (GRIIS) [Colección]. Recuperado de http://www.griis.org

GmbH Oberpfaffenhofen. (2018). Linked Data Sets [Portal]. Recuperado de https://www.app-lab.eu/linked-data-sets/

Goddard, A., Wilson, N., Cryer, P., y Yamashita, G. (2011). Data hosting infrastructure for primary biodiversity data. *BMC Bioinformatics*, 12(Suppl 15), S5. doi:10.1186/1471-2105-12-s15-s5

Google. (2018a). Google Drive [Servicio]. Recuperado de https://www.google.com/intl/es_ALL/drive/

Google. (2018b). Chrome (Windows 10/8.1/8/7 64-bit) [Navegador]. Recuperado de https://www.google.com.mx/chrome/browser/desktop/index.html

Google. (2018c). Motor de búsqueda Google [Buscador]. Recuperado de https://www.google.com.mx/

Google. (2018d). Google Hangouts [Aplicación]. Recuperado de https://hangouts.google.com/?hl=es-419

Groom, Q., Weatherdon, L., y Geijzendorffer, I. R. (2016). Is citizen science an open science in the case of biodiversity observations? *Journal of Applied Ecology*, 54(2), 612–617. doi:10.1111/1365-2664.12767

Groom, Q. J., Adriaens, T., Desmet, P., Simpson, A., De Wever, A., Bazos, I., … Vanderhoeven, S. (2017). Seven Recommendations to Make Your Invasive Alien Species Data More Useful. *Frontiers in Applied Mathematics and Statistics*, 3. doi:10.3389/fams.2017.00013

Gropp, R. E. (2016). Big, Integrative, Open Biological Data. *BioScience*, 66(4), 263–264. doi:10.1093/biosci/biw045

Gruen, N., Houghton, J., y Tooth, R. (2014). Open for business: How open data can help achieve the G20 growth target. Sydney, Australia. https://www.omidyar.com/sites/default/files/file_archive/insights/ON%20Report_061114_FNL.pdf

Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Röpert, D., Casino, A., … Triebel, D. (2017). Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database*, 2017. doi:10.1093/database/bax003

Guralnick, R. P., Hill, A. W., y Lane, M. (2007). Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters*, 10(8), 663–672. doi:10.1111/j.1461-0248.2007.01063.x

Guralnick, R., e Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics*, 25(4), 421–428. doi:10.1093/bioinformatics/btn659

Guralnick, R., Conlin, T., Deck, J., Stucky, B. J., y Cellinese, N. (2014). The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices. *PLOS ONE*, 9(12), e114069. doi:10.1371/journal.pone.0114069

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., … Page, R. (2015). Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. *ZooKeys*, 494, 133–154. doi:10.3897/zookeys.494.9352

Halpin, P., Read, A., Best, B., Hyrenbach, K., Fujioka, E., Coyne, M., … Spoerri, C. (2006). OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of

marine mammals, seabirds, and sea turtles. *Marine Ecology Progress Series*, 316, 239–246. doi:10.3354/meps316239

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, 57(2), 280–299. doi:10.1353/lib.0.0036

Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. *Digital Library Federation, Council on Library and Information Resources*, 1755 Massachusetts Ave., NW, Suite 500, Washington, DC 20036. https://www.clir.org/pubs/reports/pub91/contents/

Hoffmann, N., Berendsohn, W., Güntsch, A., Kohlbecker, A., Luther, K., y Müller, A. (2011). Biodiversity information platforms: From standards to interoperability. *ZooKeys*, 150, 71–87. doi:10.3897/zookeys.150.2166

Holetschek, J., Kelbert, P., Müller, A., Ciardelli, P., Güntsch, A., y Berendsohn, W. G. (2009). International Networking of Large Amounts of Primary Biodiversity Data. *In GI Jahrestagung* (pp. 552-564).

Holetschek, J., Dröge, G., Güntsch, A., y Berendsohn, W. G. (2012). The ABCD of primary biodiversity data access. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, 146(4), 771–779. doi:10.1080/11263504.2012.740085

Horton, T., Gofas, S., Kroh, A., Poore, G. C. B., Read, G., Rosenberg, G., … Vranken, S. (2017). Improving nomenclatural consistency: a decade of experience in the World Register of Marine Species. *European Journal of Taxonomy*, (389). doi:10.5852/ejt.2017.389

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... y Twigger, S. (2008). Big data: The future of biocuration. *Nature*, 455(7209), 47-50. doi:10.1038/455047a

Huang, X., y Qiao, G. (2011). Biodiversity databases should gain support from journals. *Trends in Ecology y Evolution*, 26(8), 377–378. doi:10.1016/j.tree.2011.05.006

Hugo, W., Hobern, D., Kõljalg, U., Tuama, É. Ó., y Saarenmaa, H. (2016). Global Infrastructures for Biodiversity Data and Services. *The GEO Handbook on Biodiversity Observation Networks*, 259–291. doi:10.1007/978-3-319-27288-7_11

ICZN. (2018). ZooBank [Colección]. Recuperado de http://www.zoobank.org

iDigBio. (2018). Integrated Digitized Biocollections [Colección]. Recuperado de https://www.idigbio.org

Innologica. (2018). Inoreader [Servicio]. Recuperado de https://www.inoreader.com/

Inoreader. (2018). Inoreader Companion (v. 4.1.5) [Complemento]. Recuperado de https://chrome.google.com/webstore/detail/inoreader-companion/kfimphpokifbjgmjflanmfeppcjimgah

International DOI Foundation. (2017). DOI® Handbook. Recuperado de: https://www.doi.org/hb.html

IODE. (2018). Ocean Biogeographic Information Systems [Colección]. Recuperado de http://www.iobis.org

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S. y Rice, P. (2013). EDAM Ontology [Ontología]. Recuperado de http://edamontology.org/page

ITIS. (2017, 8 de noviembre). Integrated Taxonomic Information System [Colección]. Recuperado de https://www.itis.gov

IUCN. (2018a). International Union for Conservation of Nature and Natural Resources [Colección]. Recuperado de http://www.iucnredlist.org

IUCN. (2018b). Global Invasive Species Database [Colección]. Recuperado de http://www.iucngisd.org

Jayasiri, S. C., Hyde, K. D., Ariyawansa, H. A., Bhat, J., Buyck, B., Cai, L., … Hidayat, I. (2015). The Faces of Fungi database: fungal names linked with morphology, phylogeny and human impacts. *Fungal Diversity*, 74(1), 3–18. doi:10.1007/s13225-015-0351-8

Johnson, N. F. (2007). Biodiversity Informatics. *Annual Review of Entomology*, 52(1), 421–438. doi:10.1146/annurev.ento.52.110405.091259

Jones, A. C. (2006). Prospects for a biodiversity grid: managing biodiversity knowledge. *Sixth IEEE International Symposium on Cluster Computing and the Grid* (CCGRID'06). doi:10.1109/ccgrid.2006.1630935

Juffe-Bignoli, D., Brooks, T. M., Butchart, S. H. M., Jenkins, R. B., Boe, K., Hoffmann, M., … Kingston, N. (2016). Assessing the Cost of Global Biodiversity and Conservation Knowledge. *PLOS ONE*, 11(8), e0160640. doi:10.1371/journal.pone.0160640

Kelbert, P., Droege, G., Barker, K., Braak, K., Cawsey, E. M., Coddington, J., … Güntsch, A. (2015). B-HIT - A Tool for Harvesting and Indexing Biodiversity Data. *PLOS ONE*, 10(11), e0142240. doi:10.1371/journal.pone.0142240

Koureas, D., Hardisty, A., Vos, R., Agosti, D., Arvanitidis, C., Bogatencov, P., … Smith, V. (2016). Unifying European Biodiversity Informatics (BioUnify). *Research Ideas and Outcomes*, 2, e7787. doi:10.3897/rio.2.e7787

Krzywinski, M. (2018). Circos Table Viewer [Software]. Recuperado de http://mkweb.bcgsc.ca/tableviewer/

Laurenne, N., Tuominen, J., Saarenmaa, H., y Hyvönen, E. (2014). Making species checklists understandable to machines – a shift from relational databases to ontologies. *Journal of Biomedical Semantic*s, 5(1), 40. doi:10.1186/2041-1480-5-40

Library of Congress. (2018). ID.LOC.GOV [Ontología]. Recuperado de http://id.loc.gov/

LSID. (s. f.). LSID Resolution Project, Life Sciences Identifier [Estándar]. Recuperado de http://www.lsid.info/

Martellos, S., y Attorre, F. (2012). New trends in biodiversity informatics. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, 146(4), 749–751. doi:10.1080/11263504.2012.740092

Mazzocchi, F. (2018). Encyclopedia of Knowledge Organization (ISKO) [Portal]. Recuperado de http://www.isko.org/cyclo/kos

Méndez-Muñoz, V., Cohen-Nabeiro, A., David, R., Ivars Camáñez, V. J., Nonell-Canals, A., Senar, M. A., … Tatoni, T. (2017). Analysis on the Graph Techniques for Data-mining and Visualization of Heterogeneous Biodiversity Data Sets. *Proceedings of the 2nd International Conference on Complexity, Future Information Systems and Risk*. doi:10.5220/0006379701440151

Michán, L. (2011). Cienciometría, información e informática en ciencias biológicas: enfoque interdisciplinario para estudiar interdisciplinas. *Ludus Vitalis*, 19(35), 239-243. http://www.ludus-vitalis.org/ojs/index.php/ludus/article/view/214/210

Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., ... y Chavan, V. (2011). Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics*, 12(15), S1. doi:10.1186/1471-2105-12-S15-S1

NCBI. (2018a). PubMed [Colección]. Recuperado de https://www.ncbi.nlm.nih.gov/pubmed/

NCBI. (2018b). NCBI Taxonomy [Colección]. Recuperado de https://www.ncbi.nlm.nih.gov/taxonomy

NCIT. (2019). National Cancer Institute Thesaurus [Tesauro]. Recuperado de: https://bioportal.bioontology.org/ontologies/NCIT?p=summary

Nelson, G., Sweeney, P., y Gilbert, E. (2018). Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. *Applications in Plant Sciences*, 6(2), e1027. doi:10.1002/aps3.1027

NLM. (2016). PubMed Tutorial [Tutorial]. Recuperado de https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/cover.html

NLM. (2019). Medical Subject Headings (MeSH) [Tesauro]. Recuperado de: https://bioportal.bioontology.org/ontologies/MESH

NSF. (2018). Open Tree of Life [Colección]. Recuperado de https://tree.opentreeoflife.org

NSU. (2018). "CRAAP" Test. Recuperado de http://nova.campusguides.com/evaluate

Page, R. D. M. (2006). Taxonomic names, metadata, and the Semantic Web. *Biodiversity Informatics*, 3(0). doi:10.17161/bi.v3i0.25

Page, R. D. M. (2008). Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, 9(5), 345–354. doi:10.1093/bib/bbn022

Page, R. D. M. (2009). bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics. *BMC Bioinformatics*, 10 Suppl 1, S5. doi:10.1186/1471-2105-10-S14-S5

Page, R. D. M. (2013). BioNames: linking taxonomy, texts, and trees. *PeerJ*, 1, e190. doi:10.7717/peerj.190

Parr, C. S., Wilson, N., Leary, P., Schulz, K., Lans, K., Walley, L., … Corrigan, Jr., R. (2014). The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal*, 2, e1079. doi:10.3897/bdj.2.e1079

Parr, C. S., Schulz, K. S., Hammock, J., Wilson, N., Leary, P., Rice, J., y Corrigan, R. J. (2016). TraitBank: Practical semantics for organism attribute data. *Semantic Web*, 7(6), 577–588. doi:10.3233/sw-150190

Parr, C. S., y Thessen, A. E. (2018). Biodiversity Informatics. *Ecological Informatics*, 375–399. doi:10.1007/978-3-319-59928-1_17

PATO. (2018). Phenotypic Quality Ontology [Ontología]. Recuperado de https://bioportal.bioontology.org/ontologies/PATO

Pennock, M. (2007). Digital Curation: A life-cycle approach to managing and preserving usable digital information. Library & Archives, 1, 34-45. http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf

Peterson, A. T., Knapp, S., Guralnick, R., Soberón, J., y Holder, M. T. (2010). The big questions for biodiversity informatics. *Systematics and Biodiversity*, 8(2), 159–168. doi:10.1080/14772001003739369

Peterson, A. T., Soberón, J., y Krishtalka, L. (2015). A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecology*, 15(1). doi:10.1186/s12898-015-00

Phenoscape. (2018). [Colección]. Recuperado de https://phenoscape.github.io

Rees, T., Vandepitte, L., Decock, W., y Vanhoorne, B. (2017). IRMNG 2006–2016: 10 years of a global taxonomic database. *Biodiversity Informatics*, 12. doi:10.17161/bi.v12i0.6522

Rees, T. (compilador). (2018). The Interim Register of Marine and Nonmarine Genera [Colección]. Recuperado de http://www.irmng.org

Reimer, L. C., Söhngen, C., Vetcininova, A., y Overmann, J. (2017). Mobilization and integration of bacterial phenotypic data—Enabling next generation biodiversity analysis through the Bac Dive metadatabase. *Journal of Biotechnology*, 261, 187–193. doi:10.1016/j.jbiotec.2017.05.004

Roberts, D., y Moritz, T. (2011). A framework for publishing primary biodiversity data. BMC Binformatics, 12(15): I1. doi:10.1186/1471-2105-12-S15-I1

Rosati, I., Bergami, C., Stanca, E., Roselli, L., Tagliolato, P., Oggioni, A., … Basset, A. (2017). A thesaurus for phytoplankton trait-based approaches: Development and applicability. *Ecological Informatics*, 42, 129–138. doi:10.1016/j.ecoinf.2017.10.014

Ruete, A. (2015). Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodiversity Data Journal*, 3, e5361. doi:10.3897/bdj.3.e5361

Santos, A. M., y Branco, M. (2012). The quality of name-based species records in databases. *Trends in Ecology & Evolution*, 27(1), 6–7. doi:10.1016/j.tree.2011.10.004

Sarr, I., Naacke, H., Bame, N., Gueye, I., y Ndiaye, S. (2014). Green and Distributed Architecture for Managing Big Data of Biodiversity. *Computing in Research and Development in Africa*, 21–39. doi:10.1007/978-3-319-08239-4_2

Sarkar, I. N. (2007). Biodiversity informatics: organizing and linking information across the spectrum of life. *Briefings in Bioinformatics*, 8(5), 347–357. doi:10.1093/bib/bbm037

Schnase, J. L., Cushing, J., Frame, M., Frondorf, A., Landis, E., Maier, D., y Silberschatz, A. (2003). Information technology challenges of biodiversity and ecosystems informatics. *Information Systems*, 28(4), 339-345.

SeaLifeBase. (2018). The SeaLifeBase Project [Blog]. Recuperado de https://sealifebaseproject.blogspot.com/2018/04/fishbase-and-sealifebase-sign-mou-with.html

Senderov, V., Georgiev, T., y Penev, L. (2016). Online direct import of specimen records into manuscripts and automatic creation of data papers from biological databases. *Research Ideas and Outcomes*, 2, e10617.

Sikes, D. S., Copas, K., Hirsch, T., Longino, J. T., y Schigel, D. (2016). On natural history collections, digitized and not: a response to Ferro and Flick. *ZooKeys*, 618, 145–158.

Silva, L. A. E., Siqueira, M. F., Pinto, F. dos S., Barros, F. S. M., Zimbrão, G., y Souza, J. M. (2016). Applying data mining techniques for spatial distribution analysis of plant species co-occurrences. *Expert Systems with Applications*, 43, 250–260. doi:10.1016/j.eswa.2015.08.031

Smith, V. S., Rycroft, S. D., Harman, K. T., Scott, B., y Roberts, D. (2009). Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. *BMC Bioinformatics*, 10(Suppl 14), S6. doi:10.1186/1471-2105-10-s14-s6

Soberón, J., y Peterson, A. T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1444), 689–698.

Suhrbier, L., Kusber, W.-H., Tschöpe, O., Güntsch, A., y Berendsohn, W. G. (2017). AnnoSys—implementation of a generic annotation system for schema-based data using the example of biodiversity collection data. *Database*, 2017. doi:10.1093/database/bax018

TDWG. (2018). Biodiversity Information Standards (TDWG) [Portal]. Recuperado de http://www.tdwg.org/

Tessarolo, G., Ladle, R., Rangel, T., y Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution*, 7(17), 6863–6870. doi:10.1002/ece3.3259

The Marine Biological Laboratory. (2018). Universal Biological Indexer and Organizer (uBio) [Colección]. Recuperado de http://www.ubio.org

The Writing Center. (2018). Literature Reviews [Tutorial]. Recuperado de https://writingcenter.unc.edu/tips-and-tools/literature-reviews/

Thessen, A., y Patterson, D. (2011). Data issues in the life sciences. *ZooKeys*, 150, 15–51. doi:10.3897/zookeys.150.1766

Tirador Ramos, J. (2010). El Dominio y su implicación para la Gestión de la Información. *Investigación bibliotecológica*, 24(50), 49-60.

TreeBASE. (2018). TreeBASE [Colección]. Recuperado de https://treebase.org

Triebel, D., Hagedorn, G., y Rambold, G. (2012). An appraisal of megascience platforms for biodiversity information. *MycoKeys*, 5, 45–63. doi:10.3897/mycokeys.5.4302

Tschöpe, O., Macklin, J. A., Morris, R. A., Suhrbier, L., y Berendsohn, W. G. (2013). Annotating biodiversity data via the Internet. *Taxon*, 62(6), 1248–1258. doi:10.12705/626.4

Turnhout, E., y Boonman-Berson, S. (2011). Databases, scaling practices, and the globalization of biodiversity. *Ecology and Society*, 16(1). doi:10.5751/ES-03981-160135

UNESCO. (2018). Tesauro de la UNESCO [Diccionario, Glosario, Tesauro]. Recuperado de http://vocabularies.unesco.org/browser/thesaurus/es/

UN Environment. (2018). Convention on International Trade in Endangered Species of Wild Fauna and Flora Trade Database [Colección]. Recuperado de https://trade.cites.org

UserLand Software. (2002). FEED Validator, RSS 2.0 SPECIFICATION. Recuperado de https://validator.w3.org/feed/docs/rss2.html#whatIsRss

VertNet. (2018). [Colección]. Recuperado de http://vertnet.org

Vos, R., Biserkov, J., Balech, B., Beard, N., Blissett, M., Brenninkmeijer, C., … Sierra, S. (2014). Enriched biodiversity data as a resource and service. *Biodiversity Data Journal*, 2, e1125. doi:10.3897/bdj.2.e1125

W3C. (1999). Web Characterization Terminology & Definitions Sheet [Portal]. Recuperado de https://www.w3.org/1999/05/WCA-terms/#Core

W3C. (2013). Web Ontology Language [Estándar]. Recuperado de https://www.w3.org/OWL/

W3C. (2014). Resource Description Framework (RDF) [Estándar]. Recuperado de https://www.w3.org/RDF/

W3C. (2015a). Semantic Web [Portal]. Recuperado de https://www.w3.org/standards/semanticweb/

W3C. (2015b). Linked Data [Estándar]. Recuperado de https://www.w3.org/standards/semanticweb/data

W3C. (2016). Extensible Markup Language (XML) [Estándar]. Recuperado de https://www.w3.org/XML/

Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., … Wooley, J. (2014). Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLOS ONE*, 9(3), e89606. doi:10.1371/journal.pone.0089606

Wang, X., Zhang, F., y Zhang, J. (2017). Biodiversity information resources. I. Species distribution, catalogue, phylogeny, and life history traits. *Biodiversity Science*, 25(11), 1223–1238. doi:10.17520/biods.2017184

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. doi:10.1371/journal.pone.0029715

Wikidata. (2019). Plug-in. Recuperado de: https://www.wikidata.org/wiki/Q184148

Willis, C., Greenberg, J., y White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology*, 63(8), 1505–1520. doi:10.1002/asi.22683

Woolcott, L. (2017). Understanding Metadata: What is Metadata, and What is it For?,. *Cataloging & Classification Quarterly*, 55(7-8), 669–670. doi:10.1080/01639374.2017.1358232

WoRMS. (2018). World Register of Marine Species [Colección]. Recuperado de http://www.marinespecies.org

Wu, L., Sun, Q., Desmeth, P., Sugawara, H., Xu, Z., McCluskey, K., … Ma, J. (2016). World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. *Nucleic Acids Research*, 45(D1), D611–D618. doi:10.1093/nar/gkw903

Xiao-Ting, D. (2012). Research Progress on Biodiversity Information System. *2012 International Conference on Computer Science and Service System*. doi:10.1109/csss.2012.81

Zeng, M. L. (2008). Knowledge organization systems (KOS). Knowledge organization, 35(2-3), 160-182. https://www.ergon-verlag.de/en/bibliotheks--informationswissenschaft/knowledge-organization-journal/knowledge-organization-double-issue-2008-2-3.php

Zermoglio, P. F., Guralnick, R. P., y Wieczorek, J. R. (2016). A Standardized Reference Data Set for Vertebrate Taxon Name Resolution. *PLOS ONE*, 11(1), e0146894. doi:10.1371/journal.pone.0146894

Anexos disponibles en línea*

*Se recomienda el uso de la última versión de los navegadores Google Chrome o Mozilla Firefox.