

INSTITUTO TECNOLÓGICO DE LA PAZ  
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN  
MAESTRÍA EN SISTEMAS COMPUTACIONALES

**MODELO DE DATOS PARA EL ANÁLISIS DE  
INFORMACIÓN GEORREFERENCIADA SOBRE  
BIODIVERSIDAD, UTILIZADO EN UN ENTORNO DE  
APLICACIONES MÓVILES**

QUE PARA OBTENER EL GRADO DE  
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:  
ALBERTO GONZÁLEZ ESPINOZA

DIRECTORES DE TESIS:  
MATI. LUIS ARMANDO CÁRDENAS FLORIDO  
DR. GUILLERMO MARTÍNEZ FLORES

LA PAZ, BAJA CALIFORNIA SUR, MÉXICO, AGOSTO 2016.



La Paz, B.C.S., **1/agosto/2016**


DEPI/345/2016


ASUNTO: Autorización de impresión

C. ALBERTO GONZÁLEZ ESPINOZA,  
ESTUDIANTE DE LA MAESTRÍA EN  
SISTEMAS COMPUTACIONALES,  
P R E S E N T E .

Con base en el dictamen de aprobación emitido por el Comité Tutorial de la Tesis denominada: "MODELO DE DATOS PARA EL ANÁLISIS DE INFORMACIÓN GEORREFERENCIADA SOBRE BIODIVERSIDAD, UTILIZADO EN UN ENTORNO DE APLICACIONES MÓVILES", mediante la opción de tesis (Proyectos de Investigación), entregado por usted para su análisis, le informamos que se AUTORIZA la impresión

ATENTAMENTE  
"CIENCIA ES VERDAD, TÉCNICA ES LIBERTAD"

  
M.C. MANUEL E. CASILLAS BROOK,  
SUBDIRECTOR ACADÉMICO.

  
INSTITUTO TECNOLÓGICO DE LA PAZ  
DIVISIÓN DE ESTUDIOS DE POSGRADO  
E INVESTIGACIÓN

c.c.p. Archivo.  
MACB/LACF/ici



## DICTAMEN DEL COMITÉ TUTORIAL

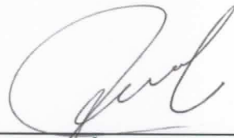
SUBDIRECCIÓN ACADÉMICA  
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN.

La Paz, B.C.S., **01 /agosto/ 2016**

**C. MC. MANUEL E. CASILLAS BROOK,**  
SUBDIRECTOR ACADÉMICO,  
P R E S E N T E.

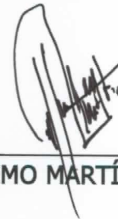
Por medio del presente, enviamos a usted dictamen del Comité Tutorial de tesis para la obtención del grado de Maestro, con los siguientes datos generales:

No. de Control M14310019	Nombre ALBERTO GONZÁLEZ ESPINOZA
Maestría en:	SISTEMAS COMPUTACIONALES
Título de la tesis: Modelo de datos para el análisis de información georreferenciada sobre biodiversidad, utilizado en un entorno de aplicaciones móviles	
<b>DICTAMEN:</b> Se autoriza el trabajo de investigación, en virtud de que realizó las correcciones correspondientes conforme a las observaciones planteadas por este Comité Tutorial.	



MC. JESÚS ANTONIO CASTRO

**A t e n t a m e n t e .**  
El Comité Tutorial



DR. GUILLREMO MARTÍNEZ FLORES



MATI. LUIS ARMANDO CÁRDENAS FLORIDO

c.c.p. Coordinador de la Maestría.  
c.c.p. Departamento de Servicios Escolares.  
c.c.p. Estudiante.

# Resumen

Las bases de datos de biodiversidad almacenan información histórica sobre la variedad de organismos que constituyen la vida sobre la Tierra. Esta información ha sido recabada por la comunidad científica durante más de 250 años de trabajo, recolectando y clasificando seres vivos alrededor de todo el planeta. Estas bases de datos guardan esencialmente la información de las especies que han sido registradas en un lugar y momento específicos. Además, cuentan con información biológica relacionada con la clasificación de organismos encontrados y datos sobre cómo se realizó el registro y quién lo hizo. Al momento de integrar toda esta información, podemos responder a cinco preguntas básicas sobre el espécimen registrado:

1. ¿A qué grupo de organismos pertenece?
2. ¿Dónde fue encontrado?
3. ¿Cuándo se encontró?
4. ¿Quién lo registró?
5. ¿De qué forma fue registrado?

Toda esta información conforma el conocimiento primario sobre la biodiversidad para la comunidad científica, pero además, al analizarse adecuadamente puede ser aplicada para la resolución de problemáticas concretas, como la designación de áreas naturales protegidas, el diseño de políticas para el manejo de la tierra y el combate de especies invasoras.

Las bodegas de datos son de gran ayuda para trabajar con esta información histórica acumulada, ya que en ellas se pueden almacenar todos los datos masivos de ocurrencias de seres vivos,

y aplicar un modelo de datos multidimensional que permita analizarla y comprenderla más fácilmente.

Los modelos de datos multidimensionales permiten analizar la información de la bodega de datos desde diferentes perspectivas, a través de formas de visualización que son fácilmente comprensibles para el usuario. Así, en una primera instancia podemos realizar la descripción de la diversidad biológica de una región, a partir del número de organismos encontrados. En este caso, sólo nos interesarían los datos relacionados con la ubicación espacial de los ejemplares. Esta sería nuestra primera dimensión. Después, podríamos añadir una segunda variable, que sería la clasificación de los organismos, con lo cual sabríamos qué cantidad de organismos de cada grupo fueron encontrados en esa región. Nuestras dimensiones ahora serían los datos espaciales y la clasificación de los organismos. Si añadimos una tercera variable, como es el tiempo, encontraremos que la información se distribuye durante diferentes momentos. Además, tanto la dimensión espacial como la del tiempo pueden visualizarse en diferentes escalas. Así, podríamos seguir agregando variables a nuestro modelo, como las organizaciones que registraron los especímenes, de que forma fueron registrados, etc. La implementación de este modelo multidimensional dentro de una bodega de datos recibe el nombre de hipercubo, y podemos utilizarlo como una herramienta para realizar consultas multidimensionales que respondan a preguntas mucho más complejas, como por ejemplo:

- ¿Qué país posee el mayor inventario de capturas de especímenes realizadas en territorio mexicano?
- ¿En qué años hubo mayor número de registros de peces en Bahía Magdalena?
- ¿Qué localidad del Noroeste de México cuenta con la mayor presencia de lobos marinos?

Además, gracias a que los datos se encuentran georreferenciados, es posible sumar dimensiones espaciales adicionales dentro del modelo, al definir polígonos de coordenadas geográficas en la base de datos, que correspondan a regiones de interés para el usuario que analiza la información. De esta forma, se pueden clasificar a todos los ejemplares dentro de una región geográfica, y analizarlos por separado del resto de la base de datos.

Finalmente, las aplicaciones para dispositivos móviles ofrecen nuevas oportunidades para desarrollar innovaciones, dentro del terreno de los sistemas de información sobre biodiversidad. Gracias a su capacidad de almacenar datos fuera de línea y a su Sistema de Posicionamiento Global (GPS, por sus siglas en inglés) incorporado, podemos acceder a la información consultada previamente en localidades fuera del alcance de Internet, así como ubicar los lugares donde se registraron los ejemplares, utilizando el sistema de geoposicionamiento del dispositivo móvil.

# Abstract

Biodiversity databases, store historical information about the variety of organisms that constitute the life on Earth. This information has been gathered by the scientific community for more than 250 years of work collecting and classifying living things around the world. These databases, essentially contain the information of the species that have been recorded in a specific place and time. In addition, they have biological information related to the classification of organisms found and details about how the record was made and who made it. At the time of integrating all of this information, we can respond to five basic questions about the registered specimen:

1. Which group of organisms it belong to?
2. Where was it found?
3. When was it found?
4. Who recorded it?
5. How was it recorded?

This information represents the primary knowledge of biodiversity for the scientific community, but also when is analyzed properly, can be applied to the resolution of specific problems, such as the designation of protected areas, the design of policies for the management of the land and combat invasive species.

Data warehouses are helpful for working with this accumulated historical information, since in them you can store all the massive data of occurrences of living beings, and apply a multidimensional data model that allows you to analyze and understand it more easily.

Multidimensional data models, allow you to analyze the information in the data warehouse from different perspectives, through forms of visualization which are easily understandable for the user. Thus, in the first instance we can make the description of the biological diversity of a region, based on the number of specimens found. In this case, only be interested us in data related to the spatial location of the specimens. This would be our first dimension. Then, we could add a second variable, that would be the classification of organisms, so we would know how many organisms of each group were found in the region. Our dimensions would now be the spatial data and the classification of organisms. If we add a third variable, as the time, we will find that the information is distributed during different times. In addition, both the spatial dimension and the time can be displayed in different scales. Thus, we could continue adding variables in our model, as the organizations that recorded the specimens, the form in which they were registered, etc. The implementation of this multidimensional model within a data warehouse, is called a hypercube, and we can use it as a tool to perform multidimensional queries that answer questions far more complex, such as:

- Which country does the largest inventory of catches of specimens made in Mexico?
- In what years was there a greater number of records of fish in Bahía Magdalena?
- What locality in the Northwest of Mexico does have the greater presence of sea lions?

In addition, because the data are geo-referenced, it is possible to add additional dimensions within the model, to define polygons of geographic coordinates in the database, corresponding to regions of interest to the user that analyzes the information. In this way, you can classify all specimens within a geographical region, and analyze them separately from the rest of the database.

Finally, applications for mobile devices offer new opportunities to develop innovations in the field of biodiversity information systems. Thanks to its ability to store data offline and its Global Positioning System (GPS) incorporated, we can access the previously consulted information being in locations outside the reach of the Internet, as well as locate the places where the specimens were registered, using the GPS of the mobile device.



# Índice general

<b>1. Introducción</b>	<b>1</b>
1.1. Antecedentes . . . . .	1
1.2. Descripción del problema . . . . .	4
1.3. Objetivos . . . . .	6
1.3.1. Objetivo general . . . . .	6
1.3.2. Objetivos específicos . . . . .	6
1.4. Justificación . . . . .	7
1.5. Limitaciones . . . . .	8
<b>2. Marco teórico</b>	<b>10</b>
2.1. Bases de datos de biodiversidad . . . . .	10
2.1.1. Datos primarios de biodiversidad . . . . .	11
2.1.2. La Infraestructura Mundial de Información de Biodiversidad . . . . .	12
2.1.3. Georreferenciación de la información . . . . .	14
2.1.4. Referenciación por nombres científicos . . . . .	20

2.1.5.	Más allá de los registros primarios: La Enciclopedia de la Vida . . . . .	21
2.2.	La inteligencia de negocios . . . . .	21
2.2.1.	Las bodegas de datos . . . . .	22
2.2.2.	Los sistemas OLAP . . . . .	23
2.2.3.	Los modelos multidimensionales . . . . .	24
2.2.4.	Las bodegas de datos y los data marts . . . . .	26
2.2.5.	Enfoques para la construcción de las bodega de datos . . . . .	27
2.2.6.	Arquitectura de la bodega de datos . . . . .	27
2.2.7.	Modelado lógico de una bodega de datos . . . . .	28
2.2.8.	Las consultas multidimensionales . . . . .	31
2.2.9.	SQL Server Analysis Services . . . . .	31
2.3.	Componentes del sistema de consulta móvil . . . . .	32
2.3.1.	Las aplicaciones móviles . . . . .	32
2.3.2.	El formato JSON . . . . .	33
2.3.3.	El contenido algorítmico . . . . .	35
<b>3.</b>	<b>Desarrollo del modelo de datos</b>	<b>38</b>
3.1.	Extracción, transformación y carga . . . . .	39
3.1.1.	Extracción de la información . . . . .	39
3.1.2.	Transformación de la información . . . . .	42
3.1.3.	Carga de la información . . . . .	54

3.2. Diseño del modelo multidimensional . . . . .	55
3.2.1. La tabla de hechos . . . . .	56
3.2.2. Definición de métricas . . . . .	57
3.2.3. Diseño lógico del modelo de datos . . . . .	58
3.2.4. Definición de las dimensiones y sus jerarquías . . . . .	59
3.2.5. Prueba de las dimensiones . . . . .	65
3.3. Diseño de las consultas multidimensionales . . . . .	66
3.3.1. Una consulta en dos dimensiones . . . . .	67
3.3.2. Rotando el cubo . . . . .	68
3.3.3. Descendiendo en la jerarquía . . . . .	69
3.3.4. Una consulta en tres dimensiones . . . . .	71
3.3.5. Añadiendo la dimensión del tiempo . . . . .	73
<b>4. Construcción del cliente OLAP móvil</b>	<b>75</b>
4.1. Consulta y visualización de un mapa sobre biodiversidad . . . . .	76
4.2. Servicio web para la consulta del hipercubo . . . . .	80
4.3. Interfaz de consulta multidimensional . . . . .	81
<b>5. Resultados</b>	<b>84</b>
5.1. Funcionalidades del cliente OLAP móvil . . . . .	86
5.2. Preguntas contestadas . . . . .	87

5.2.1. ¿Qué país posee el mayor inventario de capturas de especímenes realizadas en territorio mexicano? . . . . .	88
5.2.2. ¿A qué bases de datos podemos acudir para obtener la información sobre las familias de plantas con flores de la región? . . . . .	89
5.2.3. ¿En qué años hubo mayor número de registros de peces en Bahía Magdalena?	90
5.2.4. ¿Qué región del Noroeste de México cuenta con la mayor presencia de lobos marinos? . . . . .	91
5.2.5. ¿Cómo ha variado la forma de registrar la biodiversidad de la región a los largo del tiempo? . . . . .	92
<b>6. Conclusiones y trabajo futuro</b>	<b>94</b>
<b>A. Instrucciones SQL para la inserción de las regiones</b>	<b>97</b>
<b>Bibliografía</b>	<b>100</b>

# Capítulo 1

## Introducción

### 1.1. Antecedentes

Durante las pasadas dos décadas se han realizado diversos esfuerzos por parte de diferentes organismos internacionales, encaminados a concretar sistemas de información sobre la biodiversidad de nuestro planeta. Entre los retos a los que se han tenido que enfrentar estos sistemas informáticos de biodiversidad se encuentra la dispersión de la información, muchas veces almacenada en las bases de datos de las colecciones biológicas de diferentes instituciones de investigación y museos de historia natural, alrededor del mundo, así como la dificultad para integrar índices estandarizados de nombres científicos para las especies.

En el año de 1999, la Organización para la Cooperación y el Desarrollo Económico (OCDE por sus siglas en inglés), creó la Infraestructura Mundial de Información sobre la Biodiversidad <sup>1</sup> (GBIF por sus siglas en inglés), con el objetivo de coordinar y poner en línea todas las bases de datos electrónicas para diferentes grupos de organismos.

A lo largo de estos años, este proyecto ha logrado integrar una base de datos centralizada, con información referente a las ocurrencias de especies registradas por la comunidad científica alrededor del mundo. Cada uno de los registros de las ocurrencias se encuentra georeferenciado

---

<sup>1</sup><http://www.gbif.org/>

y ofrece información complementaria referente al lugar, el momento y la forma de captura y registro de los ejemplares. Toda la información acumulada es accesible a través de los servicios web que brinda este sistema. Hasta el momento de escribir este documento, existe un total de 515'702,963 registros de ocurrencias de especímenes, pertenecientes a 1'454,694 especies, dentro de 13,818 catálogos de datos.

Dentro de nuestro país, la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad<sup>2</sup> (CONABIO) se ha convertido en el principal proveedor de información sobre las especies del territorio mexicano, que alimenta las bases de datos de la GBIF, al tener a su disposición una gran base de datos de las colecciones biológicas de las instituciones con las cuales colabora.

Uno de los grandes logros alcanzados, al disponer de la información provista por GBIF, ha sido el desarrollo de un gran número de proyectos que explotan esta base de datos centralizada, alimentando diferentes aplicaciones de información sobre biodiversidad. Podemos acceder a la información de cada uno de estos desarrollos en el propio portal de GBIF, (sección Using Data<sup>3</sup>). De estas aplicaciones destacan tres, por su relación con este proyecto:

- La Enciclopedia de la Vida<sup>4</sup> (EOL por sus siglas en ingles), creada en en 2008 con el propósito de proveer un catálogo en línea con la información de cada especie sobre la Tierra. Este sistema brinda información validada proveniente de diversas fuentes reconocidas alrededor del mundo, como museos de historia natural, sociedades filantrópicas, expertos científicos y otras dentro de una base de datos masiva, presentada en un portal de fácil acceso para su consulta en Internet. Este sistema ofrece un servicio de búsqueda de especies por nombre común o científico, y como salida final despliega una página con la información recabada en su base de datos sobre cada especie obtenida en la búsqueda. Entre la información mostrada se encuentran los nombres comunes en distintos idiomas, colecciones fotográficas, su árbol evolutivo, su descripción general, las referencias bibliográficas de los trabajos donde las estudian y por supuesto, los mapas con información georreferenciada provista por la GBIF. Además, cuenta con un conjunto de servicios web para que las aplicaciones creadas por otros puedan obtener esta misma información sobre

---

<sup>2</sup><http://www.gbif.org/publisher/ff90b050-c256-11db-b71b-b8a03c50a862>

<sup>3</sup><http://www.gbif.org/usingdata/dataapplications>

<sup>4</sup><http://eol.org/>

las especies indexadas.

- El Mapa de la Vida<sup>5</sup> (MOL por sus siglas en inglés), creado en el año 2012 con el propósito de brindar a sus usuarios información geográfica sobre la biodiversidad del planeta, a través de los mapas interactivos disponibles en su portal de Internet. Este sistema proporciona datos sobre el área geográfica en la que se distribuyen las especies en su hábitat y los puntos geográficos específicos en donde los investigadores han registrados organismos. Existen dos formas en las que los usuarios pueden acceder a esta información. La primera consiste en buscar las especies por su nombre científico, obteniendo una lista de la cual se selecciona la especie a visualizar en el mapa. La segunda forma consiste en elegir un punto sobre el mapa para la búsqueda. El sistema encontrará las especies que se localizan alrededor de este punto (100 km a la redonda), mostrando un listado similar al de la búsqueda antes descrita, para elegir que especie se desea visualizar. El sistema no permite visualizar diferentes especies en un mismo mapa, y se limita solo a cinco grupos generales de organismos: anfibios, aves, mamíferos, plantas y peces. Este sistema cuenta además con una aplicación para dispositivos móviles que permite realizar las mismas consultas que soporta la aplicación web, aunque con una presentación mucho más simplificada.
- Animals + plants es una aplicación para teléfonos móviles, creada con el apoyo del Ministerio Federal de Educación e Investigación de Alemania, que permite a los usuarios de dispositivos Android subir observaciones de fauna georreferenciadas. Los datos extraídos de GBIF ayudan a proporcionar una lista de las especies que los usuarios pueden esperar encontrar en una localidad en particular, y el proyecto publica los datos de las observaciones a través del portal de GBIF.

Además de ser utilizados por estos proyectos de información biológica, los datos proporcionados por el GBIF son fuente de datos en artículos científicos, así como en proyectos encaminados a resolver problemáticas concretas, como el combate de especies invasoras y la designación de áreas naturales protegidas.

Con el creciente auge de los dispositivos móviles entre los usuarios de Internet, se abre la posibilidad de aprovechar estas fuentes de información para crear aplicaciones que aprovechen las

---

<sup>5</sup><http://www.mol.org/>

ventajas para el trabajo en campo que proporcionan los teléfonos inteligentes, que implementen funciones de consulta y que faciliten a los usuarios conocer la información sobre la biodiversidad en regiones de su interés.

Estos sistemas de información constituyen las bases sobre las cuales se desarrolló la planificación de este proyecto.

## 1.2. Descripción del problema

Investigadores, estudiantes y aficionados de la biodiversidad se enfrentan a la problemática de que la información sobre las especies aún permanece dispersa, difícil de acceder o de descifrar (Balke et al., 2013). Como un ejemplo de las dificultades de acceso a la información, podemos ver que los sistemas de información sobre biodiversidad antes descritos tienen limitaciones en las funciones de consulta. La dispersión de la información se refiere a que aunque se ha logrado recolectar gran cantidad de información, los datos relevantes de biodiversidad (registros georreferenciados, clasificación taxonómica, referencias bibliográficas, descripción del inventario físico de las colecciones, datos multimedia, etc.) no se encuentran en una única fuente, sino que debemos extraerla de diferentes orígenes. Sobre la dificultad para descifrar los datos, podemos ver que la consulta de información a través de datos técnicos como los nombres científicos, permite obtener resultados más precisos, pero requiere de un nivel de conocimientos que solo poseen los expertos en la materia.

Un desafío central para la comunidad informática que aborda temas sobre la biodiversidad, es proporcionar los medios para compartir y sintetizar rápidamente los datos y los conocimientos disponibles para construir un mapa unificado global de la biodiversidad fácilmente accesible. Dicho mapa habría de proporcionar datos en bruto y la información resumida sobre la biodiversidad y su cambio en todo el mundo en múltiples escalas (Guralnick and Hill, 2009).

Las capas de este mapa no se limitarían a mostrar los datos en bruto, sino que también podrían presentar la información compilada, tales como riqueza estimada de especies, características de comportamiento y morfológicas, o la proliferación de linajes que contienen fenotipos o genotipos



únicos (Guralnick and Hill, 2009).

Al igual que en la mayoría de las ciencias biológicas y de la Tierra, en los sistemas de información de biodiversidad y ecología la ubicación es muy importante. Gran parte de los datos de la biodiversidad y los ecosistemas están *georreferenciados* (referidos a una ubicación geográfica única) y también se distinguen por ser *referenciados a través de las especies*. Los datos genéticos se asocian frecuentemente con una especie o subespecie, en la investigación de invasiones y extinciones se realiza un seguimiento a nivel de especie, y mucha de la caracterización de un ecosistema se describe a través del número y la distribución de sus especies constituyentes (Schnase et al., 2007). Los usuarios de los sistemas sobre biodiversidad requieren consultar esta información, lo cual implica acceder a los sistemas que la contienen en su mayor parte.

Los usuarios de los datos e información sobre biodiversidad y ecosistemas, incluyendo los administradores del uso de suelo, políticos, educadores, organizaciones no gubernamentales y la industria, necesitan técnicas de visualización para una mejor comprensión de los datos y las relaciones entre éstos, los procesos naturales y la gestión de acciones en el tiempo. Esto brinda oportunidades para la investigación en técnicas avanzadas de despliegue, incluyendo la visualización de incertidumbre, del uso adaptativo y de datos multidimensionales (Schnase et al., 2007).

Además, debido al trabajo de campo que realizan los investigadores y estudiantes, la solución informática que se proponga para consultar y visualizar la información sobre biodiversidad deberá responder a las necesidades de movilidad de este grupo particular de usuarios, brindando la posibilidad de acceder a la información consultada estando fuera del alcance de las redes de Internet.

Los dispositivos de cómputo móvil, utilizados por un gran número de usuarios en la actualidad, permiten, desde el punto de vista de requerimientos de hardware, que un desarrollo de esta índole sea posible sin necesitar de una mayor inversión. De esta forma, podemos centrar la descripción de la problemática en el diseño del sistema de software involucrado.

Este sistema de software deberá ser capaz de acceder a la información de las especies contenida en las fuentes antes descritas, a través de los servicios proporcionados por la GBIF y la EOL.

Además, se deben aprovechar las posibilidades de combinar el GPS de los dispositivos móviles con los datos georreferenciados de las bases de datos, ya que de esta forma se puede conseguir un sistema que permita acceder directamente a la información, bastando con la posición geográfica obtenida por el dispositivo móvil para iniciar una consulta.

Finalmente, en virtud del gran volumen de la información colectada, es necesario el diseño de un sistema capaz de optimizar el tiempo de respuesta, al procesar la gran cantidad de datos que se obtendrán desde distintas fuentes.

## **1.3. Objetivos**

### **1.3.1. Objetivo general**

Diseñar un modelo de datos multidimensional que permita consultar y analizar la información histórica sobre las ocurrencias de organismos a nivel regional, que pueda ser accedida fuera de línea desde un dispositivo móvil.

### **1.3.2. Objetivos específicos**

1. Diseñar un modelo de datos multidimensional que permita a los usuarios analizar la información desde diferentes dimensiones y jerarquías, incluyendo una dimensión espacial para clasificar los datos en regiones geográficas definidas por el usuario.
2. Construir una bodega de datos que implemente este modelo multidimensional, utilizando un fragmento de la base de datos centralizada de la GBIF, para probar el modelo diseñado.
3. Desarrollar una aplicación móvil que permita visualizar los resultados de las consultas multidimensionales, y que además utilice el GPS del dispositivo móvil para acceder directamente a la información de los organismos, localizada en la GBIF y la EOL.
4. Brindar acceso fuera de línea a la información de biodiversidad consultada previamente.

## 1.4. Justificación

La diversidad biológica —o biodiversidad—, nos provee de aire limpio, agua limpia, alimentos, ropa, refugio, medicinas y el disfrute estético. La biodiversidad y los ecosistemas que la soportan, contribuyen con billones de dólares a las economías nacionales y mundiales, directamente a través de industrias como la agricultura, la silvicultura, la pesca y el ecoturismo e indirectamente a través de servicios biológicamente mediados tales como la polinización de plantas, la dispersión de semillas, tierras de pastoreo, eliminación de dióxido de carbono, fijación de nitrógeno, el control de inundaciones, ruptura de residuos y el biocontrol de plagas de los cultivos. La riqueza biológica de los ecosistemas es uno de los más importantes factores que influyen en la estabilidad y la salud de nuestro medio ambiente (Schnase et al., 2007).

La biodiversidad se distribuye por toda la Tierra, con la mayor concentración en las regiones tropicales, especialmente en los países en desarrollo y en los océanos. En contraste, la información científica sobre la biodiversidad está en gran parte concentrada en los grandes centros de los países desarrollados, sobre todo en las colecciones científicas de museos de historia natural, herbarios y repositorios de microorganismos. En la actualidad, es más probable, por ejemplo, que la información sobre las plantas de una región particular de África se almacene en un herbario en Europa, que en su país de origen (Edwards et al., 2000). Este dato es especialmente relevante para nuestro país, que es considerado a nivel mundial un lugar de gran riqueza en este campo (Rands et al., 2010).

Por lo anterior, los sistemas de información sobre la biodiversidad cobran especial importancia tanto para los investigadores que buscan ampliar los conocimientos sobre la variedad y distribución de las distintas formas de vida, como para las organizaciones a cargo de la toma de decisiones sobre el manejo y cuidado de los recursos naturales. En este contexto, los sistemas informáticos sobre biodiversidad juegan un papel importante, ya que brindan información relevante para el desarrollo de investigaciones, al mismo tiempo que dan sustento a la toma de decisiones que contribuyan a la protección del medio ambiente, así como para el aprovechamiento de los recursos naturales.

Por otro lado, los dispositivos de cómputo móvil ofrecen la posibilidad de explotar la informa-

ción en situaciones donde las computadoras personales no están disponibles, al mismo tiempo que poseen características particulares de gran utilidad para el trabajo en campo. Tal y como nos indica Hardisty et al. (2013): “los avances en las comunicaciones móviles ofrecen numerosas oportunidades para la innovación. Smartphones y Tablet-PC con la ubicación GPS incluida pueden ser utilizados fácilmente en el campo, creando oportunidades tanto para los servicios innovadores de recopilación de datos y la información del usuario. También son particularmente innovadoras para hacer referencia a productos, tales como las claves de identificación. Aplicaciones como éstas pueden ser utilizadas para generar observaciones etiquetadas por su localización y acompañadas de imágenes probatorias, para ser subidas posteriormente a una base de datos central”.

En nuestro país, son muchas las instituciones de investigación que se benefician con el acceso a las bases de datos centralizadas de la biodiversidad mundial que poseen la GBIF y la EOL dentro de su infraestructura, —especialmente las instituciones dedicadas a la investigación biológica y a la conservación—. Solo dentro de nuestro estado, Baja California Sur, encontramos tres instituciones de reconocimiento nacional e internacional que desarrollan investigación en estos campos, como son: el Centro Interdisciplinario de Ciencias Marinas (CICIMAR), el Centro de Investigaciones Biológicas del Noroeste (CIBNor) y la Universidad Autónoma de Baja California Sur (UABCS).

Además, en el mercado actual de las aplicaciones móviles aún no se ha desarrollado un sistema con las características antes mencionadas, por lo que se abre una oportunidad importante para incursionar en el campo de la informática para la biodiversidad, y lograr un impacto relevante dentro de esta área.

## 1.5. Limitaciones

Este proyecto estará limitado al uso de la información sobre biodiversidad contenida en sistemas ya desarrollados, concretamente en la GBIF y la EOL, ya que la combinación de los servicios web de ambos ofrece la información mínima que se requiere para resolver las necesidades de información detectadas durante el análisis documental.

De esta forma, la cantidad de información estará limitada a los datos que la comunidad científica ponga a disposición de estos dos sistemas, lo cual también representa una limitante en cuanto a la actualización de la información, ya que los datos recolectados pueden tardar un tiempo considerable en ser liberados. Por lo tanto, la información para el análisis provendrá en su mayoría de datos históricos.

Debido al factor tiempo, el desarrollo de la solución de software se limitó a la plataforma de aplicaciones móviles Android.

Esta aplicación se alimentará de bases de datos globales de información, por lo que su utilización no estará restringida a una región o país, así que la interfaz de la aplicación móvil se limitará al idioma inglés, con el propósito de llegar al mayor número de usuarios posibles, dejando abierta la posibilidad de incluir otros idiomas, en versiones posteriores.

Por la misma razón de que se utilizarán datos globales de información, hay que considerar las limitantes de hardware y de comunicación en diferentes entornos donde se pueda emplear este sistema, debido principalmente al gran volumen de los datos involucrados, por lo que es necesario implementar mecanismos de filtrado o discriminación de los datos, a fin de lograr que el tiempo de respuesta sea aceptable.

# Capítulo 2

## Marco teórico

### 2.1. Bases de datos de biodiversidad

La biodiversidad se define como toda variación de la base hereditaria en todos los niveles de organización, desde los genes en una población local o especie, hasta las especies que componen toda o una parte de una comunidad local, y finalmente las comunidades que componen la parte viviente de los múltiples ecosistemas del mundo. Abarca, por tanto, todos los tipos y niveles de variación biológica (Nuñez et al., 2003). Otra definición, mucho más compacta y sencilla de manejar, afirma que la biodiversidad es la variedad de genes, especies y ecosistemas que constituyen la vida sobre la Tierra (Rands et al., 2010).

De acuerdo con Date (2001), una base de datos es un sistema informático de registros guardados en tablas, con un arreglo establecido, con base en un propósito que ordena, mantiene, procesa, presenta, recupera y genera información con las siguientes ventajas:

1. Acumulan una gran cantidad de información en poco espacio.
2. Sistematizan los datos de acuerdo con los objetivos, metas y necesidades del proyecto.
3. Proporcionan un eficiente acceso a la información.
4. Permiten realizar búsquedas a partir de diferentes criterios, ya sean simples o combinados.

5. Pueden procesar los datos de forma cualitativa o cuantitativa.
6. Interrelacionan la información combinando o cruzando variables.
7. Permiten actualizar la información de forma fácil y rápida.
8. Pueden ser compatibles entre sí, es decir, puede existir interoperabilidad entre ellas.
9. Efectúan diversos cálculos por medio de funciones incorporadas y extraen los datos por medio de consultas.
10. Se pueden extender o ampliar por medio de módulos o tablas relacionales.
11. Actualmente se pueden diseñar, almacenar, manejar y analizar incluso en dispositivos móviles.

Para poder lograr un control estructurado del almacenamiento físico de los datos, independiente de su estructura lógica, que permita modificar, extraer y garantizar su integridad son necesarios una serie de procedimientos que ejecuten esta labor; a esta serie de procedimientos (software), se le conoce como Sistema de Gestión de Base de Datos (SGBD) (Vaisman and Zimányi, 2014).

Una base de datos de biodiversidad, por lo tanto, es aquella que almacena registros que contienen información sobre diferentes organismos que fueron localizados por los investigadores en distintos lugares y momentos.

La información biológica que contienen estas bases de datos se puede separar en tres niveles de organización: molecular/genética, de especies y de ecosistemas (Lane et al., 2007). Este proyecto, se enfocará en la información a nivel especie, por lo que es necesario comprender algunos de los términos relacionados con este tipo de información, así como tener un visión general del proceso de clasificación de las especies.

### **2.1.1. Datos primarios de biodiversidad**

Mucha de la investigación ecológica se basa en Datos Primarios de Biodiversidad (PBD por sus siglas en inglés), también llamados Registros Primarios de Biodiversidad (PBR, por sus siglas en

inglés) cuando estos forman parte de una base de datos. Un PBD es una pieza de información que detalla una ocurrencia: el avistamiento o muestreo de un individuo perteneciente a una especie en un momento y lugar específicos. Es decir, un PBD describe (en su forma más básica) lo que ha sido observado o recolectado, y dónde y cuándo fue. Otros datos adicionales pueden mejorar este triplete básico: cuanto mayor acceso tenemos a la información primaria de biodiversidad, mejores conclusiones podemos obtener de estos registros en términos de fiabilidad (Otegui et al., 2013).

En los estudios de ámbito local, un investigador puede obtener directamente muestras de datos, ya sea en el campo o en museos y herbarios. Sin embargo, en los estudios de ámbito global, esto se convierte en una labor abrumadora. Por lo tanto, los facilitadores de datos se vuelven extremadamente útiles para la recopilación de datos para los estudios a nivel global. Un facilitador de datos es una iniciativa (institución, base de datos o proyecto) que vincula diferentes fuentes de datos en un marco común, con el fin de facilitar el acceso a los datos de todo el conjunto a través de una única puerta de enlace (Otegui et al., 2013).

### **2.1.2. La Infraestructura Mundial de Información de Biodiversidad**

La Infraestructura Mundial de Información de Biodiversidad (GBIF por sus siglas en inglés) es actualmente la mayor iniciativa de este tipo. La GBIF fue propuesta por la OCDE dentro del “Foro de Mega Ciencia” (ahora Foro Mundial de Ciencia) en el año 1999, y se estableció oficialmente por los gobiernos en 2001 con el objetivo de “hacer que los datos primarios de la biodiversidad del mundo estén libre y universalmente disponibles a través de Internet”. Por medio de una red mundial de 57 países y 47 organizaciones, la GBIF “promueve y facilita la movilización, el acceso, el descubrimiento y uso de la información sobre la presencia de organismos a través del tiempo y en todo el planeta”. Técnicamente, este facilitador trabaja, entre otras cosas, como un agregador de información sobre la biodiversidad y, en el momento de su almacenamiento, permite el acceso a más de 317 millones de PBR puestos a disposición por 342 instituciones, en lo sucesivo conocidos como proveedores de datos, a partir de un portal de información común (<http://data.gbif.org/>). La GBIF tiene su sede en Copenhague, Dinamarca, y tiene una estructura descentralizada con nodos nacionales y regionales (Otegui et al.,



2013).

Los proveedores de datos son la parte central de la GBIF. Se componen, entre otros, por centros de investigación, universidades o redes de información sobre biodiversidad, quienes mantienen actualizada una colección de datos y la ponen a disposición del público (Otegui et al., 2013).

Cada proveedor comparte uno o más conjuntos de datos (datasets), también llamados recursos de datos, que contiene registros individuales - los PBD reales - y metadatos, por ejemplo, de pertenencia de la información a una institución, derechos de propiedad intelectual, información de una campaña de muestreo o datos relativos a un grupo taxonómico particular. Los PBD son compartidos usando un estándar de datos normalizados (Darwin Core, DWC) al cual se asignan los campos de la base de datos. Los metadatos sobre los recursos son publicados por el portal de información de la GBIF mediante uno de varios mecanismos de intercambio de recursos, por ejemplo, mediante la recolección por medio de los protocolos de comunicación DiGIR o TAPIR o a través de la publicación directa utilizando el Integrated Portal Toolkit (IPT) en el Catálogo de Metadatos de la GBIF (<http://metadata.gbif.org/catalogue/>)(Otegui et al., 2013).

Los esfuerzos de la GBIF están enfocados sobre los PBD de los especímenes y a nivel de especie porque estos datos, a diferencia de la mayoría de los datos genéticos/moleculares y ecológicos, aún no se encuentran digitalizados. Sin embargo, este tipo de PBD son críticamente importantes para la sociedad, la ciencia y la sustentabilidad del futuro (Lane et al., 2007).

De acuerdo a Lane et al. (2007), los tipos de datos y servicios de la GBIF y sus participantes están encaminados a proveer a través de Internet:

- Datos georreferenciados de especímenes (ver fig. 2.1)
- un índice electrónico de nombres científicos y
- un medio para enlazar en un solo conjunto datos de fuentes distintas (p. e. secuencias de ADN, imágenes de especies, caracteres morfológicos, observaciones de especies y datos sobre ecosistemas) para responder a preguntas complejas.

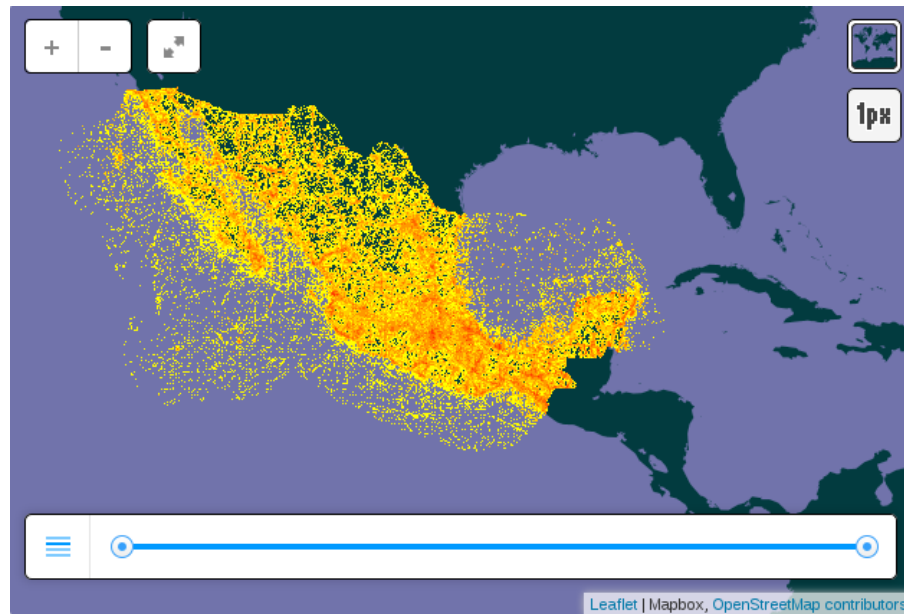


Figura 2.1: Mapa de los Registros Primarios de Biodiversidad recolectados dentro del territorio mexicano, obtenido de <http://www.gbif.org/country/MX/summary>. En este aparece representada cada ocurrencia de un ejemplar con un punto de color amarillo. Conforme el número de ocurrencias en un mismo sitio va aumentando, el color de éste tiende a un naranja más intenso.

### 2.1.3. Georreferenciación de la información

La georreferenciación es el proceso de relacionar la información con una ubicación geográfica, utilizando un sistema de ubicaciones geospaciales (coordenadas geográficas tales como la longitud y la latitud) (Hill, 2009).

Uno de los aspectos fundamentales de cualquier registro de biodiversidad es la descripción de la ubicación de la ocurrencia de una especie, ya sea la descripción textual del lugar o en un formato más fácilmente analizable como la latitud y la longitud (Hill et al., 2009).

#### 2.1.3.1. *Datum* geodésico

Estos sistemas de coordenadas geospaciales, forman parte de un sistema de referencia conocido como *datum* geodésico o sistema geodésico. Los *datums* geodésicos definen el tamaño y forma de la Tierra, así como el origen y orientación de los sistemas de coordenadas usados por los mapas.

Cientos de *datums* han sido empleados para formular descripciones posicionales, desde que la primera estimación del tamaño de la Tierra fuera realizada por Aristóteles. Los *datums* han evolucionado desde aquellos que describían una Tierra esférica, hasta los modelos elipsoidales derivados de años de mediciones satelitales (Dana, 1995).

Los sistemas geodésicos tienen dos componentes básicos: el elipsoide de referencia y el geoide.

Como se muestra en la fig. 2.2, un elipsoide viene definido por dos parámetros: el semieje mayor y el semieje menor. En el caso de la Tierra estos se corresponderían con el radio ecuatorial y el radio polar respectivamente. La relación existente entre estas dos medidas define el grado de achatamiento del elipsoide. En particular, se establece un factor de achatamiento según

$$f = \frac{r_1 - r_2}{r_1} \quad (2.1)$$

siendo  $r_1$  el semieje mayor y  $r_2$  el semieje menor (Olaya, 2014).

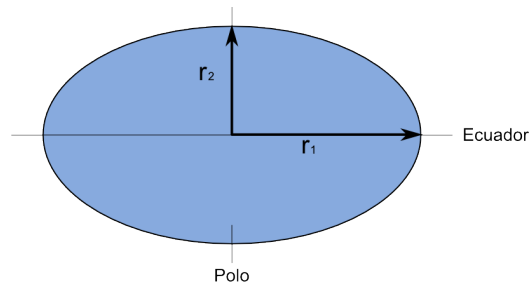


Figura 2.2: Parámetros que definen el elipsoide, obtenido de Olaya (2014)

El elipsoide es la forma geométrica que mejor se adapta a la forma real de la Tierra, y por tanto la que mejor permite idealizar esta, logrando un mayor ajuste (Olaya, 2014).

Actualmente, los modelos elipsoidales de la Tierra son utilizados por tener una precisión extendida y un cálculo de orientación sobre grandes distancias (Dana, 1995). Por ejemplo, el elipsoide WGS-84 es muy empleado en la actualidad, pues es el utilizado por los GPS (Olaya, 2014).

El geoide es la otra superficie de referencia, definida como la superficie tridimensional en cuyos puntos la atracción gravitatoria es constante. Se trata de una superficie equipotencial que resulta de suponer los océanos en reposo y a un nivel medio (el nivel es en realidad variable como

consecuencia de las mareas, corrientes y otros fenómenos) y prolongar estos por debajo de la superficie terrestre. La particularidad del geoide reside en que en todos sus puntos la dirección de la gravedad es perpendicular a su superficie (Olaya, 2014).

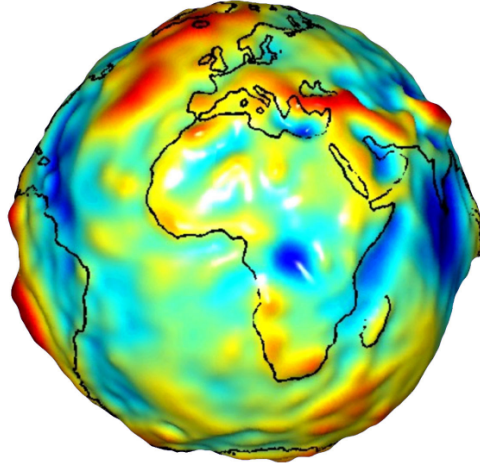


Figura 2.3: Representación gráfica del geoide, obtenida de Olaya (2014)

El geoide no es, sin embargo, una superficie regular como el elipsoide, y presenta protuberancias y depresiones que lo diferencian, como puede observarse en la fig. 2.3. La densidad de la Tierra no es constante en todos sus puntos, y ello da lugar a que el geoide sea una superficie irregular como consecuencia de las anomalías gravimétricas que dichas variaciones de densidad ocasionan (Olaya, 2014).

Lógicamente, el elipsoide, por su naturaleza más simple, no puede recoger toda la variabilidad del geoide, por lo que estas dos superficies presentan diferencias, cuyo máximo es generalmente del orden de  $\pm 100$  metros. Estas diferencias se conocen como alturas geoidales (Olaya, 2014).

Al igual que en el caso de los elipsoides, existen diversos geoides de referencia, y estos no son constantes en el tiempo sino que evolucionan para adaptarse a las modificaciones que tienen lugar sobre la superficie terrestre (Olaya, 2014).

Cuando se trabaja con un elipsoide general, este se sitúa de tal modo que tanto la posición de su centro de gravedad como su plano ecuatorial coincidan con los terrestres. Por el contrario, cuando el elipsoide es local, estas propiedades no han de cumplirse necesariamente, y el elipsoide a solas resulta insuficiente ya que carecemos de información sobre su posicionamiento con respecto a la

superficie terrestre (Olaya, 2014).

Surge así el concepto de *datum*, que es el conjunto formado por una superficie de referencia (el elipsoide) y un punto en el cual “enlazar” este al geoide. Este punto se denomina punto astronómico fundamental (para su cálculo se emplean métodos astronómicos), o simplemente punto fundamental, y en él el elipsoide es tangente al geoide. La altura geoidal en este punto es, como cabe esperar, igual a cero. La vertical al geoide y al elipsoide son idénticas en el punto fundamental (Olaya, 2014).

Para un mismo elipsoide pueden utilizarse distintos puntos fundamentales, que darán lugar a distintos *datum* y a distintas coordenadas para un mismo punto (Olaya, 2014).

### 2.1.3.2. Sistemas de coordenadas geográficas

Disponiendo de un modelo preciso para definir la forma de la Tierra, podemos establecer ya un sistema para codificar cada una de las posiciones sobre su superficie y asignar a estas las correspondientes coordenadas. Puesto que la superficie de referencia que consideramos es un elipsoide, lo más lógico es recurrir a los elementos de la geometría esférica y utilizar estos para definir el sistema de referencia. De ellos derivan los conceptos de latitud y longitud, empleados para establecer las coordenadas geográficas de un punto (Olaya, 2014).

De acuerdo a Olaya (2014) el sistema de coordenadas geográficas es un sistema de coordenadas esféricas mediante el cual un punto se localiza con dos valores angulares:

- La latitud  $\phi$  es el ángulo entre la línea que une el centro de la esfera con un punto de su superficie y el plano ecuatorial (ver fig. 2.4). Las líneas formadas por puntos de la misma latitud se denominan paralelos y forman círculos concéntricos paralelos al ecuador. Por definición, la latitud es de  $0^\circ$  en el ecuador que divide al globo en los hemisferios norte y sur. La latitud puede expresarse especificando si el punto se sitúa al norte o al sur, por ejemplo  $24^\circ, 21' 11''$  N, o bien utilizando un signo, en cuyo caso los puntos al Sur del ecuador tienen signo negativo.
- La longitud  $\lambda$  es el ángulo formado entre dos de los planos que contienen a la línea de

los polos (ver fig. 2.4). El primero es un plano arbitrario que se toma como referencia y el segundo es el que, además de contener a la línea de los polos, contiene al punto en cuestión. Las líneas formadas por puntos de igual longitud se denominan meridianos y convergen en los polos. Como meridiano de referencia internacional se toma el que pasa por el observatorio de Greenwich, en el Reino Unido. Este divide a su vez el globo en dos hemisferios: el este y el oeste. La longitud puede expresarse especificando si el punto se sitúa al este o al oeste, por ejemplo  $32^{\circ}, 12' 43''$  E, o bien utilizando un signo, en cuyo caso los puntos al Oeste del meridiano de referencia tienen signo negativo.

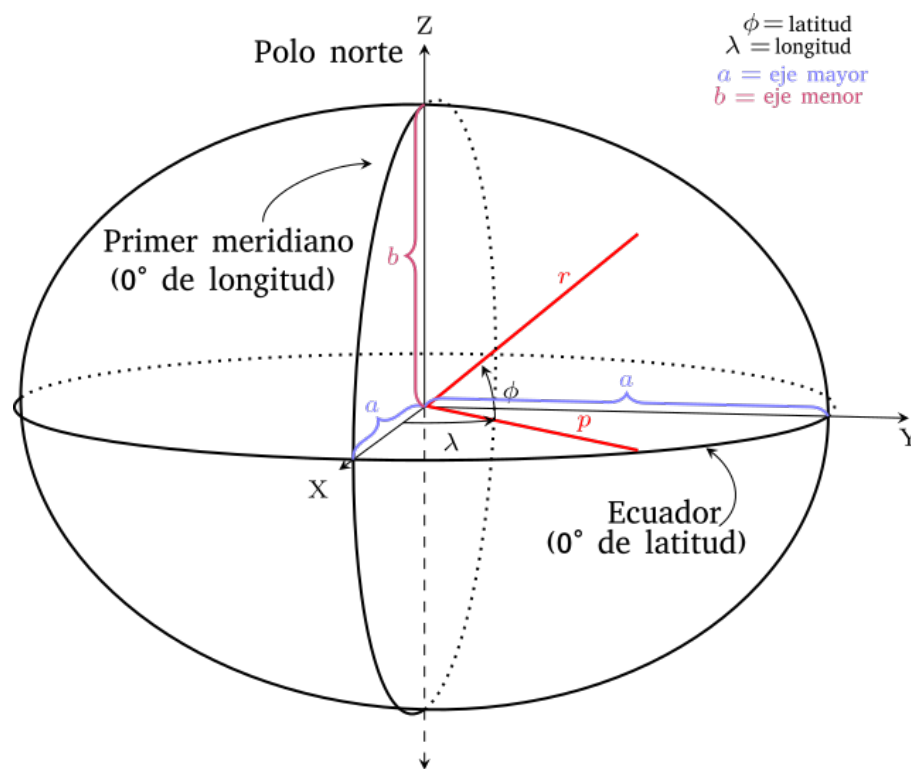


Figura 2.4: Representación gráfica del elipsoide de las coordenadas geodésicas. En este se aprecian los ángulos que forman la latitud ( $\phi$ ) y la longitud ( $\lambda$ ). Para primera se forma un ángulo desde el plano ecuatorial hasta la dirección vertical de una línea normal (señalada por  $r$ ) que pasaría por el punto buscado, mientras que para la segunda, el ángulo es formado desde el meridiano de referencia (o Primer meridiano) hasta otro meridiano (señalado por  $p$ ) que pasaría por el punto señalado; al converger ambos ángulos, se posicionaría geográficamente el punto especificado. Modificado de Dana (1995).

### 2.1.3.3. El datum WGS-84

Como se mencionó antes, existen cientos de *datums* geodésicos alrededor del mundo, y uno de los más utilizados es el World Geodetic System 1984 (WGS-84). Se trata de un sistema de referencia basado en un modelo elipsoidal, con un sistema de coordenadas geodésicas de latitud, longitud y elevación, que utiliza una definición aproximada del nivel del mar, para calcular la altitud (Dana, 1995).

En la GBIF se utiliza este sistema para georreferenciar los datos sobre biodiversidad.

### 2.1.3.4. Métodos de georreferenciación

De acuerdo a Wieczorek et al. (2004) hay varios métodos con los cuales la descripción de un lugar puede ser georreferenciada, entre los que encontramos:

**Método de punto** El más utilizado es el método de “Punto”, con el cual un solo par de coordenadas es asignado a cada ubicación.

**Método de figuras** El método de figuras es un método conceptualmente simple que delinea una localidad utilizando uno o más polígonos, almacenando puntos y polilíneas. La combinación de estas formas puede representar una ciudad, parque, río, cruce, o cualquier otra característica o combinación de características que se encuentran en un mapa. Aunque es simple de describir, la tarea de generar estas formas puede ser difícil. La creación de figuras es poco práctica sin la ayuda de mapas digitales, software de Sistemas de Información Geográfica (SIG), y la experiencia, todo lo cual puede ser relativamente caro. Además, el almacenamiento de una figura en una base de datos es considerablemente más complicado que el almacenamiento de un solo par de coordenadas.

**Método de cuadros delimitadores** Una forma común para describir una característica geográfica es utilizar un cuadro delimitador de dos pares de coordenadas que juntos forman un rectángulo (en la proyección adecuada) que abarca la localidad que se describe. El método

de cuadros delimitadores es una variación limitada del método de figuras, por el cual sólo los puntos o los rectángulos proyectados pueden ser descritos.

**Método de punto y radio** El método de punto-radio describe una localidad como un par de coordenadas y una distancia desde ese punto (es decir, un círculo), la combinación de los cuales abarca la descripción plena de la localidad.

Los datos de las ocurrencias que encontramos en la GBIF pueden ser consultados a través de ubicaciones espaciales, ya sea proporcionando al servicio un solo par de coordenadas de un sitio (método de punto) o a través de un conjunto de puntos que forman un polígono sobre el mapa (método de figuras).

#### 2.1.4. Referenciación por nombres científicos

Otra de las tareas de la GBIF es generar un catálogo de nombres científicos de todos los organismos conocidos, el cual es un elemento requerido para poder realizar minería de datos a través de los tres niveles de organización biológica (genético/molecular, de especies y de ecosistemas) en una sola consulta (Lane et al., 2007).

Esta nomenclatura científica se basa en la clasificación taxonómica de Carlos Linneo (1707-1778), la cual utiliza nombres binomiales en latín, en donde la primera parte del nombre (género) es compartido por un grupo de organismos similares, mientras que la segunda parte, el epíteto, diferencia a los miembros de ese grupo (p. e. el género *Quercus* cuenta con alrededor de 600 especies). Una clasificación jerárquica similar es seguida por el género que está agrupado por familias, familias dentro de órdenes, órdenes dentro de clases y clases dentro de phylum (Hardisty et al., 2013).

Actualmente es posible hacer consultas sobre los PBR a base de texto a través de los servicios Web de la GBIF, utilizando los nombres científicos de las más de 1.4 millones de especies catalogadas en su base de datos.



### 2.1.5. Más allá de los registros primarios: La Enciclopedia de la Vida

Para complementar la información contenida en los PBR provistos por la GBIF, en este proyecto se utilizará la información biológica complementaria de los servicios Web de la Enciclopedia de la Vida (EOL por sus siglas en inglés).

La EOL trabaja para ser una proveedora libre de fuentes comprensibles y autorizadas de información sobre cada forma de vida en la Tierra. La EOL es una colección de artículos, páginas web enlazadas, imágenes, sonidos, vídeos, etc., organizados para dar un enfoque integral a la información. Fue creada en 2008 ([http://eol.org/info/the\\_history\\_of\\_eol](http://eol.org/info/the_history_of_eol)) y actualmente reúne 1'349,944 páginas de Taxones dedicadas a la información de los organismos o grupos de organismos, 1'808,786 imágenes, 68,100 participantes y 251 socios de contenido que suministran información para el sitio web (EOL.org). La EOL ofrece contenido en varios idiomas con una visión general de los organismos, complementada con información a profundidad, recursos multimedia, mapas donde el organismo puede ser hallado, nombres comunes y científicos de los organismos, enlaces a recursos externos o sitios de sus socios, las referencias a la información proporcionada y datos sobre la de actividad reciente asociada con cada página (Rucker, 2014).

Los servicios Web de consulta provistos por la EOL, soportan la búsqueda por texto a través tanto de los nombres científicos de los organismos, como de los nombres comunes en varios idiomas.

## 2.2. La inteligencia de negocios

La inteligencia de negocios se compone de una colección de metodologías, procesos, arquitecturas y tecnologías que transforman datos en bruto en información manejable y útil para la toma de decisiones de las organizaciones. La inteligencia de negocios y los sistemas de soporte a la toma de decisiones proveen asistencia para manejar varios niveles organizacionales para el análisis estratégico de la información. Estos sistemas recolectan cantidades enormes de datos y los reducen a una forma que puede ser utilizada para analizar el comportamiento organizacional. Esta transformación de los datos está integrada por una serie de tareas que toman los datos

de origen y, a través de un proceso de extracción, transformación, integración y limpieza, los almacenan en un repositorio común llamado bodega de datos (Data Warehouse). Estas bodegas de datos han sido desarrolladas como una parte integral de los sistemas de apoyo a la toma de decisiones, para proveer a los usuarios una infraestructura que les permita obtener respuestas eficientes y precisas a preguntas complejas (Vihervaara et al., 2010).

A pesar de ser usualmente relacionadas con el entorno empresarial y de negocios, las bodegas de datos pueden ser utilizadas para facilitar el análisis de la información de todo tipo de organizaciones, incluyendo el sector gubernamental y de investigación científica, por lo que representan una importante herramienta para el análisis de la información relacionada con este proyecto. De esta forma, no es difícil encontrar un símil entre el funcionamiento de una bodega de datos empresarial, con una de biodiversidad: en ambas se almacena una gran cantidad de datos históricos, los cuales describen las operaciones de una organización (en el caso de la biodiversidad, esta organización estaría conformada por los distintos grupos de investigadores que recolectan los datos sobre los organismos), y cómo se ha comportado esta organización con el paso del tiempo (en el caso de las ocurrencias de organismos, podríamos observar cómo se distribuyen los especímenes en una región a través del tiempo, cómo ha variado la forma de registrarlos, qué países han acumulado la mayor cantidad de registros sobre una especie en particular, etc.).

### **2.2.1. Las bodegas de datos**

Una bodega de datos es un repositorio de datos integrados obtenidos de varias fuentes para el propósito específico del análisis de datos multidimensional. Más específicamente, se define como una colección de datos orientados a temas, integrados, no volátiles y de tiempo variable, para soportar la toma de decisiones (Vaisman and Zimányi, 2014).

Los sistemas tradicionales de Procesamiento de Transacciones en Línea (OLTP, por sus siglas en inglés) son inapropiados para apoyar a la toma de decisiones, y las redes de computadoras de alta velocidad, por sí solas, no resuelven el problema de accesibilidad a la información (Jarke et al., 2002). Algunos ejemplos de sistemas OLTP son los sistemas de administración de recursos humanos, de asignación de créditos bancarios, de recuperación y control de cartera o de control

de seguros, etc.

La función principal de los sistemas OLTP es dar soporte a las necesidades diarias de la empresa u organización, siendo sistemas normalmente optimizados para el manejo de un conjunto predefinido de transacciones. Así, mientras los sistemas OLTP almacenan sólo información reciente, las bodegas de datos cuentan con datos históricos y resumidos. Además, las bodegas de datos permiten integrar datos de fuentes heterogéneas y posibilitan el Procesamiento Analítico en Línea (OLAP, por sus siglas en inglés) (Jarke et al., 2002). OLAP, es un paradigma distinto a OLTP, y está orientado específicamente a analizar los datos en las bases de datos organizacionales para dar soporte a la toma de decisiones, enfocándose en las consultas, particularmente, en las consultas analíticas (Vaisman and Zimányi, 2014).

Dada la complejidad de las relaciones entre las entidades involucradas, las consultas del OLAP requieren múltiples operaciones de cruces y agregaciones sobre relaciones normalizadas, sobrecargando de esta forma la base de datos normalizada (Jarke et al., 2002). Las técnicas de indexado aplicadas en OLTP no son eficientes en este caso: se necesitan de nuevas técnicas de indexado y optimización para las consultas OLAP (Vaisman and Zimányi, 2014).

De esta forma, la necesidad de un modelo de bases de datos diferente para soportar el OLAP llevó a la noción de las bodegas de datos, las cuales son (usualmente) grandes repositorios de datos consolidados de diferentes fuentes (internos y externos a la organización), que son actualizados fuera de línea y siguen un modelo multidimensional. Estando dedicadas al análisis de bases de datos, las bodegas de datos pueden estar diseñadas y optimizadas para soportar eficientemente las consultas OLAP. Adicionalmente, las bodegas de datos son también usadas para soportar otros tipos de tareas de análisis, como reportes, minería de datos y análisis estadístico (Vaisman and Zimányi, 2014).

### **2.2.2. Los sistemas OLAP**

Los sistemas de Procesamiento Analítico en Línea (o sistemas OLAP) se encargan de todas las tareas para acceder, analizar y explotar la información contenida en una bodega de datos. (Vaisman and Zimányi, 2014)

De acuerdo a Jarke et al. (2002), las operaciones típicas realizadas por los clientes OLAP incluyen:

- *Roll up* (incremento del nivel de agregación).
- *Drill down* (decremento del nivel de agregación).
- *Slice* (selección y proyección).
- *Pivo* (reorientación de la vista multidimensional).

Más allá de estas operaciones OLAP básicas, es posible tener otras aplicaciones clientes dentro las bodegas de datos tales como:

- Herramientas para reportes y consultas.
- Sistemas de Información Geográfica (GIS, por sus siglas en inglés).
- Minería de datos (encontrar patrones y tendencias en la bodega de datos).
- Sistemas de apoyo a la toma de decisiones (DSS, por sus siglas en inglés).
- Sistemas de información ejecutivas (EIS, por sus siglas en inglés).
- Estadísticas.

Las aplicaciones OLAP ofrecen al usuario una vista multidimensional de los datos, lo cual es un tanto diferente del enfoque relacional tradicional; de esta forma sus operaciones necesitan de un soporte especial y personalizado. Este soporte es dado por los sistemas multidimensionales de bases de datos y los servidores relacionales OLAP (Jarke et al., 2002).

### 2.2.3. Los modelos multidimensionales

Las bodegas de datos y los sistemas OLAP están basados en el modelo multidimensional, en el cual se visualizan los datos en un espacio  $n$ -dimensional, usualmente llamado un cubo de datos o hipercubo (Vaisman and Zimányi, 2014).

Un modelo multidimensional de datos, consiste en una representación de la información como un conjunto de hechos enlazados con varias dimensiones. Un hecho, representa el punto central de nuestro análisis, y típicamente incluye atributos llamados métricas o medidas. Las métricas son normalmente valores numéricos que permiten una evaluación cuantitativa de varios aspectos de una organización (Vaisman and Zimányi, 2014).

Por ejemplo, en este proyecto una métrica fundamental es el número de organismos que fueron encontrados en un punto geográfico particular. Otras métricas serían el número de sitios donde se encontraron ejemplares, la cantidad de especies distintas detectadas en una localidad, etc.

Las dimensiones son usadas para visualizar las métricas desde diversas perspectivas. Las dimensiones también tienen atributos asociados que las describen. En la fig. 2.5 se representa un cubo de datos con tres dimensiones, en donde los valores de los atributos de las dimensiones, corresponden a los valores de los ejes del cubo, mientras que los indicadores son las métricas que obtenemos al combinar dos o más dimensiones (Vaisman and Zimányi, 2014).

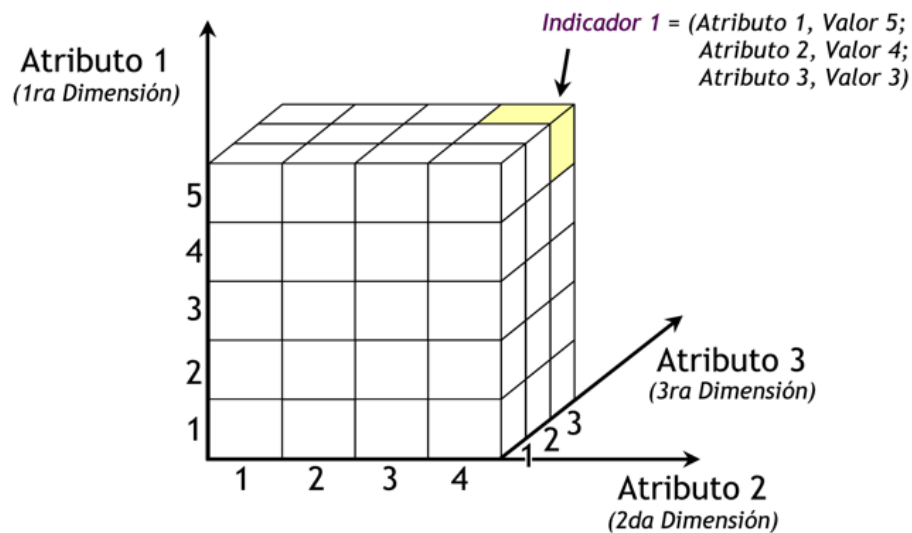


Figura 2.5: Representación de un cubo de datos o hipercubo, en donde las variables asociadas existen a lo largo de varios ejes o dimensiones. La intersección de las mismas representa la medida, indicador o el hecho que se está evaluando. En este cubo existen tres dimensiones, cada una con sus respectivos valores asociados. Obtenida de Aguilar Yelpiz (2012).

Por ejemplo, entre las dimensiones de la información sobre biodiversidad tendríamos la clasificación taxonómica de las especies, la descripción del lugar de la ocurrencia, el momento en

que ocurrió, el tipo de registro, entre otras. En este caso, encontraríamos diferentes atributos entre estas dimensiones. Por ejemplo, en la dimensión de clasificación taxonómica tendríamos el nombre de la especie, el identificador numérico de la misma, el género al que pertenece, etc.

Un cubo de datos puede ser sencillo o denso dependiendo de si tiene métricas asociadas con cada combinación de los valores dimensionales (Vaisman and Zimányi, 2014).

Un nivel dimensional representa la granularidad, o nivel de detalle en el cual las métricas están representadas por cada dimensión del cubo. Las dimensiones además, normalmente incluyen jerarquías que permiten a los usuarios explorar las métricas en varios niveles de detalle. Ésto se logra al definir una secuencia de asignaciones relacionando el nivel bajo, donde se encuentran los niveles detallados, con el nivel alto, donde se encuentran los conceptos generales (Vaisman and Zimányi, 2014).

Por ejemplo, podemos encontrar diferentes jerarquías en las dimensiones mencionadas antes, como la de clasificación de las especies, donde el nivel más detallado sería la especie y ascenderíamos en la jerarquía taxonómica pasando por distintos niveles (género, familia, etc.) hasta llegar a los reinos, el nivel de clasificación más general. Otras dimensiones donde encontramos claras jerarquías son los lugares (coordenadas geográficas, nombre de la localidad, municipio, estado, país) y el momento de la ocurrencia (día, mes, año).

Una agregación de métricas toma lugar cuando una jerarquía es recorrida (Vaisman and Zimányi, 2014). Por ejemplo, al moverse en una jerarquía desde el nivel mes hasta el nivel año, se producirán valores agregados de las ocurrencias de organismos por varios años.

#### **2.2.4. Las bodegas de datos y los data marts**

Una bodega de datos está orientada al análisis de los datos de una organización entera. En ocasiones, existen casos en los que departamentos o divisiones particulares solo requieren una porción de la bodega de datos organizacional especializada para sus necesidades. Estas bodegas de datos departamentales son llamadas data marts. Sin embargo, estos data marts no son necesariamente exclusivos de un departamento; ellos pueden ser compartidos con otras partes

interesadas dentro de la organización (Vaisman and Zimányi, 2014).

### 2.2.5. Enfoques para la construcción de las bodega de datos

De acuerdo a Vaisman and Zimányi (2014) existen dos enfoques que podemos seguir para construir una bodega de datos:

- Bottom-up, en el que la bodega de datos es vista como una colección de data marts. En este enfoque, la bodega de datos se construye primero desarrollando los data marts más pequeños y entonces se les une para obtener la bodega de datos completa.
- Top-down, que es el enfoque clásico, en el que la bodega de datos es concebida como un repositorio centralizado para la organización entera. En este enfoque, un data mart es a veces sólo una vista lógica de la bodega de datos.

### 2.2.6. Arquitectura de la bodega de datos

De acuerdo a Vaisman and Zimányi (2014), la arquitectura de una bodega de datos consiste en varias capas:

- La capa oculta. Compuesta por la extracción, transformación y carga (ETL, por sus siglas en inglés), usada para proveer datos al interior de la bodega de datos desde las bases de datos operacionales y otras fuentes de datos que pueden ser internas o externas a la organización, y un área de puesta en escena, la cual es una base de datos intermedia donde toda la integración y transformación de los datos es ejecutada antes de cargar los datos dentro de la bodega de datos.
- La capa de la bodega de datos. Compuesta por la bodega de datos empresarial y/o varios data marts y repositorios de metadatos, almacenando información acerca de la bodega de datos y su contenido.

- La capa OLAP. Compuesta de un servidor OLAP, el cual provee una vista multidimensional de los datos, sin tener en cuenta la forma real en la cual los datos están almacenados en el sistema subyacente.
- La capa frontal. Usada para analizar y visualizar los datos. Esta contiene las herramientas para el cliente tales como utilidades OLAP, herramientas de reporte, herramientas estadísticas y herramientas de minería de datos.

### 2.2.7. Modelado lógico de una bodega de datos

De acuerdo a Vaisman and Zimányi (2014), hay varios enfoques para implementar un modelo multidimensional, dependiendo de la forma en la que el cubo de datos es almacenado. Estos enfoques son:

- OLAP Relacional (ROLAP por sus siglas en inglés). Almacena los datos en bases de datos relacionales, y soporta extensiones de SQL y métodos especiales de acceso, para implementar eficientemente el modelo de datos multidimensional y las operaciones relacionadas.
- OLAP Multidimensional (MOLAP por sus siglas en inglés). Almacena los datos en estructuras de datos multidimensionales especializadas (p. e., arreglos) e implementa las operaciones OLAP sobre esas estructuras.
- OLAP Híbrido (HOLAP por sus siglas en inglés). Combina ambos enfoques.

En los sistemas ROLAP los datos multidimensionales son almacenados en tablas relacionales. Además, para incrementar el desempeño, las agregaciones son procesadas previamente en tablas relacionales. Estas agregaciones, junto con las estructuras de indexado, ocupan un gran espacio de la base de datos. Por otra parte, como los datos multidimensionales residen en tablas relacionales, las operaciones OLAP deben ser realizadas en tales tablas, resultando en sentencias SQL usualmente complejas. Finalmente, en los sistemas ROLAP, todo el manejo de los datos depende de un SGBD subyacente. Esto tiene varias ventajas debido a que las bases de datos relacionales están bien estandarizadas y proveen gran capacidad de almacenamiento (Vaisman and Zimányi, 2014).



En los sistemas MOLAP los cubos son almacenados en arreglos multidimensionales, combinados con técnicas de indexado y hashing. Por lo tanto, las operaciones OLAP pueden ser implementadas eficientemente, ya que tales procesos son muy naturales y simples de implementar. El manejo de datos en MOLAP es realizado por un motor multidimensional, el cual generalmente provee menos espacio de almacenamiento que los sistemas ROLAP. Normalmente, las estructuras típicas de índices (p. e., árboles-B, o árboles-R) son usados para indexar dimensiones pequeñas (p. e., una dimensión de productos o tiendas), y las dimensiones densas (como la dimensión del tiempo) son almacenadas en listas de arreglos multidimensionales. Cada nodo hoja del árbol de índices apunta a tales arreglos, proveyendo consultas y almacenamiento eficientes en el cubo, debido a que los índices por lo general se crean en memoria. Normalmente, los sistemas MOLAP son usados en bodegas de datos con un número de dimensiones relativamente pequeño. Para datos de alto nivel dimensional, se utilizan sistemas ROLAP. Finalmente, los sistemas MOLAP son propietarios, lo cual reduce la portabilidad (Vaisman and Zimányi, 2014).

Mientras que los sistemas MOLAP ofrecen menos capacidad de almacenamiento que los ROLAP, ellos proporcionan un mejor desempeño cuando los datos multidimensionales son consultados o agregados. Así, los sistemas HOLAP se benefician de la capacidad de almacenamiento de ROLAP y de las capacidades de procesamiento de MOLAP. Por ejemplo, un servidor HOLAP puede almacenar grandes volúmenes de información detallada en una base de datos relacional, mientras que las agregaciones son mantenidas en un almacenamiento MOLAP separado. (Vaisman and Zimányi, 2014).

Las herramientas OLAP actuales soportan una combinación de los modelos anteriores. Sin embargo, muchos de estas herramientas dependen de una bodega de datos subyacente implementada en un sistema de gestión de bases de datos relacionales. Es por esta razón que en el diseño de las dimensiones para la construcción de las bodegas de datos se siguen los lineamientos del modelado de bases de datos relacionales, a través de distintos esquemas o representaciones gráficas, conocidos como estrella, copo de nieve y constelación (Vaisman and Zimányi, 2014).

### **2.2.7.1. Esquema en estrella**

En esta representación del modelo multidimensional tenemos una tabla de hechos central, y un conjunto de tablas de dimensiones, con una tabla para cada dimensión (Vaisman and Zimányi, 2014).

En el esquema en estrella encontramos que las tablas de dimensiones están, en general, no normalizadas. Por lo tanto, contienen datos redundantes, especialmente para representar las jerarquías (Vaisman and Zimányi, 2014).

Por otro lado, la tabla de hechos está usualmente normalizada: sus llaves son la unión de las llaves foráneas debido a que su unión funcionalmente determina todas las métricas, mientras que no hay dependencia funcional entre los atributos de las llaves foráneas (Vaisman and Zimányi, 2014).

### **2.2.7.2. Esquema en copo de nieve**

Un esquema en copo de nieve evita las redundancias de los esquemas en estrella normalizando las tablas de las dimensiones. Por lo tanto, una dimensión es representada por varias tablas relacionadas por restricciones de integridad referencial. Adicionalmente, como en el caso del esquema en estrella, las restricciones de integridad referencial también relacionan la tabla de hechos y las tablas de las dimensiones en el mayor nivel de detalle (Vaisman and Zimányi, 2014).

### **2.2.7.3. Esquema en constelación**

Un esquema en copo de nieve es una combinación de los esquemas en estrella y copo de nieve, donde algunas dimensiones están normalizadas y otras no (Vaisman and Zimányi, 2014).

### 2.2.8. Las consultas multidimensionales

Así como SQL es un lenguaje para la manipulación de bases de datos relacionales, MDX (Multi-Dimensional eXpressions) es un lenguaje para la definición y consulta de bases de datos multidimensionales. Aunque a primera vista parece que MDX se asemeja a SQL, son significativamente diferentes uno del otro. Mientras que SQL opera sobre tablas, atributos y tuplas, MDX trabaja sobre cubos de datos, dimensiones, jerarquías y miembros (en el nivel de la instancia). MDX es un estándar de facto soportado por muchos proveedores de herramientas OLAP (Vaisman and Zimányi, 2014).

### 2.2.9. SQL Server Analysis Services

De acuerdo a Vaisman and Zimányi (2014), Microsoft SQL Server provee un conjunto de herramientas para la inteligencia de negocios, a través de una plataforma integrada para construir aplicaciones analíticas, compuesta por tres componentes principales, descritos a continuación:

- Analysis Services es una herramienta OLAP que provee capacidades analíticas y de minería de datos. Es usado para definir, consultar, actualizar y manejar bases de datos OLAP. En este sistema el lenguaje MDX es utilizado para obtener información. Los usuarios pueden trabajar con datos OLAP a través de herramientas tipo cliente (Excel u otros clientes OLAP) que interactúan con el componente servidor de Analysis Services.
- Integration Services soporta los procesos ETL, los cuales son usados para cargar y actualizar las bodegas de datos periódicamente. Integration Services es utilizado para extraer datos desde varias fuentes de datos; para combinar, limpiar y resumir estos datos; y, finalmente, para llenar la bodega de datos con los datos resultantes.
- Reporting Services es usado para definir, generar, guardar y manejar reportes. Los reportes pueden ser construidos a partir de varios tipos fuentes, incluyendo las bodegas de datos y los cubos OLAP. Los reportes pueden ser personalizados y distribuidos en una variedad de formatos. Los usuarios pueden visualizar los reportes con varios clientes, tales como navegadores de Internet u otros tipos de clientes para reportes.

## 2.3. Componentes del sistema de consulta móvil

Dado que el hipercubo construido con el modelo multidimensional diseñado será utilizado dentro de un entorno de aplicaciones móviles, es necesario conocer las principales características del entorno seleccionado, en este caso la plataforma Android. Además, el medio de comunicación utilizado serán los servicios Web en formato JSON, por lo que también es necesario describir la estructura de este formato. Por último, dadas las limitaciones de hardware de los dispositivos móviles y los requerimientos particulares del sistema desarrollado, es necesario utilizar contenido algorítmico para mejorar la velocidad de acceso a los datos desde la aplicación móvil.

### 2.3.1. Las aplicaciones móviles

Los dispositivos móviles y sus aplicaciones son herramientas importantes para acceder a la información cuando las computadoras de escritorio no están disponibles (Do et al., 2015). Por ejemplo, en un estudio de 4,125 usuarios de dispositivos móviles en el 2011 encontró que un usuario móvil promedio invirtió aproximadamente 59.23 minutos por día en sus dispositivos móviles, y la duración media de una sesión con una aplicación es de aproximadamente 71.56 segundos (Böhmer et al., 2011). Un informe realizado por Gartner (2013) prevé que para el año 2017, aproximadamente el 86 % de los dispositivos vendidos en todo el mundo va a estar ejecutando uno de los cuatro principales sistemas operativos móviles: Android, iOS, Windows Phone y BlackBerry. Las aplicaciones ejecutadas sobre estos dispositivos pueden acceder a diferentes datos provistos por su hardware, por ejemplo la información de geolocalización (Do et al., 2015).

Este proyecto se basará en el desarrollo de una aplicación de cómputo móvil para el sistema operativo Android, debido a las facilidades que brinda para programar aplicaciones utilizando herramientas de desarrollo de código abierto multiplataforma, así como por tratarse del sistema operativo móvil de mayor utilización en la actualidad.

### 2.3.1.1. El sistema operativo Android

Android OS es un sistema operativo de código abierto que utiliza un sistema basado en permisos para ejecutar las aplicaciones, junto con una estructura de separación de programas en ejecución (sandboxing) con el fin de reforzar la seguridad. Todas las aplicaciones que requieren acceso a cualquier recurso (por ejemplo, la lectura de los contactos del dispositivo o la grabación de audio a través del micrófono) deben solicitar los permisos correspondientes durante la instalación. Estos permisos se definen dentro de un archivo de manifiesto en el paquete instalador de aplicaciones, llamado el archivo “AndroidManifest.xml”. Un usuario puede permitir el acceso a la aplicación a todos los recursos que ha indicado que se requieren, o bien rechazar la instalación de la aplicación (Do et al., 2015).

Android utiliza la máquina virtual Dalvik para ejecutar aplicaciones (y la capa de aplicación y servicios middleware) que están escritas en Java. Estas aplicaciones se almacenan en archivos comprimidos llamados Archivos de Paquetes Android (APK por sus siglas en inglés). Con el fin de ejecutar una aplicación, la máquina virtual Dalvik lee y ejecuta el archivo “classes.dex”, contenido en el archivo APK de la aplicación, que contiene el código ejecutable Dalvik. Cada aplicación también se ejecuta en su propia máquina virtual Dalvik, con el fin de mejorar la seguridad (este proceso es conocido como sandboxing). Otros archivos contenidos en el archivo APK incluyen el archivo de manifiesto - que contiene información como las declaraciones de los recursos que la aplicación requiere y la actividad principal que realizará al ejecutarse -, y el archivo “resources.arsc” - que contiene algunos de los recursos (tales como cadenas en diferentes idiomas) de la aplicación en un formato binario comprimido - (Do et al., 2015).

### 2.3.2. El formato JSON

Tanto los servicios Web de la GBIF como de la EOL, utilizan el formato JavaScript Object Notation (JSON) para la transmisión de los datos que son consultados.

JSON es un formato de intercambio de datos ligero basado en texto independiente del lenguaje, derivado de los objetos literales del lenguaje de programación estándar ECMAScript (JavaScript), fácil de escribir y leer para los seres humanos. Tiene un formato de datos que es intercambiable

con las estructuras de datos incorporadas por los lenguajes de programación, lo que elimina el tiempo de traducción y reduce la complejidad y tiempo de procesamiento. JSON está construido con solo dos estructuras de datos: una colección de parejas nombre-valor y una lista ordenada de valores (Abd El-Aziz and Kannan, 2014).

De acuerdo a Abd El-Aziz and Kannan (2014), estas dos estructuras se especifican de la siguiente manera:

Un objeto es un conjunto desordenado de parejas nombre/valor. Su forma externa es una cadena envuelta en dos llaves con dos puntos entre los nombres y valores, y una coma entre los valores y nombres. Por ejemplo:

```
{
    "name" : "Omar",
    "properties" : {
        "height" : 1.70,
        "age" : 20,
        "designation" : "Doctor"
    }
}
```

Un arreglo, por otro lado, es una colección ordenada de valores. Un arreglo comienza con un corchete de apertura ([) y termina con un corchete de cierre (]). Por ejemplo:

```
[ "Monday", "Tuesday", "Wednesday", "Thursday", "Friday" ]

[
    [0, -1, 0],
    [1, 0, 0],
    [0, 0, 1]
]
```

Un valor puede ser una cadena de caracteres entre comillas dobles, un número (real o flotante), un valor booleano (true o false), un valor nulo (null), un objeto o un arreglo. Una cadena es una secuencia de cero o más caracteres Unicode, envuelta en comillas dobles, con escapes de

diagonal invertida. Un carácter se representa como una sola cadena de caracteres. Un número se puede representar como entero, real, o de punto flotante. JSON no admite octal o hexadecimal, porque es mínimo. No tiene valores de NaN o infinito, porque no quiere estar atado a ninguna representación interna particular. Los números no llevan comillas.

### 2.3.3. El contenido algorítmico

Entre las limitaciones del hardware de los dispositivos móviles, encontramos que en la mayoría de los casos se cuenta con una cantidad de memoria RAM justa para su funcionamiento, y un limitado espacio de almacenamiento físico. Además, debido a que la información debe poder ser consultada fuera de línea, se hace uso de archivos temporales (o caché) para almacenarla al final de cada consulta, de forma que después de suspender o cerrar la aplicación, al volver a ejecutarla ésta información esté disponible sin la necesidad de acceder a Internet de nueva cuenta.

De esta forma se hace más rápida la carga de los datos y se puede consultar los datos fuera de línea, tal y como se busca en los objetivos de este proyecto. Sin embargo, debido a que la cantidad de datos se incrementa rápidamente, resulta necesario utilizar estrategias para mejorar la velocidad de respuesta del sistema, para lo cual hacemos uso de estructuras de datos y contenido algorítmico. Así, se utilizaron estructuras de datos como una manera de almacenar y organizar la información para facilitar su acceso y modificación (Cormen et al., 2009). Para este desarrollo en particular, se implementaron listas ligadas ordenadas para almacenar los conjuntos de datos. Sobre éstas estructuras se aplicó el algoritmo de búsqueda binaria, para hacer más rápido el acceso a los datos.

Informalmente, un algoritmo es un procedimiento computacional bien definido que recibe un valor, o un conjunto de valores, como entrada y produce algún valor, o conjunto de valores, como salida. Un algoritmo es, entonces, una secuencia de pasos computacionales que transforman las entradas en salidas (Cormen et al., 2009).

También podemos ver un algoritmo como una herramienta para resolver un problema computacional bien definido. La definición del problema especifica en términos generales la relación

deseada entre las entradas y salidas. El algoritmo, entonces, describe un procedimiento computacional específico para lograr esa relación entrada/salida (Cormen et al., 2009).

Aunque el uso de algoritmos está implícito al momento de desarrollar cualquier código de un programa, decimos que nuestra aplicación utiliza contenido algorítmico cuando ésta requiere resolver un problema computacional particular, que sea inherente a su funcionamiento, y que requiera de un procedimiento computacional para obtener el resultado deseado.

Dos de las clases más comunes de algoritmos en la computación clásica son los de ordenación y búsqueda. Dentro de estos últimos se encuentra el de búsqueda binaria.

### 2.3.3.1. La búsqueda binaria

En las ciencias de la computación, un algoritmo de búsqueda binaria o de intervalo medio, busca la posición de un valor objetivo dentro de un arreglo ordenado. El algoritmo de búsqueda binaria puede ser clasificado como un algoritmo dicotómico de búsqueda divide y vencerás, ejecutándose en tiempo logarítmico (Cormen et al., 2009).

Este algoritmo parte del hecho de que, si un arreglo está ordenado, podemos revisar el punto medio de la secuencia contra el valor buscado y, en caso de no encontrar concordancia, eliminar la mitad de la secuencia para la siguiente iteración, en donde se repite este procedimiento teniendo el tamaño de la porción restante de la secuencia en cada momento, mientras no se llegue a un punto de parada (ya sea encontrar el valor buscado, o ya no poder seguir repitiendo el procedimiento de división, lo cual significa que el valor no se encuentra en la secuencia). De esta forma, se logra un algoritmo que en su peor caso se ejecuta en  $O(\lg n)$  comparaciones, donde  $n$  es el número de elementos del arreglo de entrada del problema,  $\lg$  es el logaritmo binario, y  $O$  es una constante que engloba al costo de las operaciones complementarias al procedimiento de búsqueda en sí (Cormen et al., 2009).

Por ejemplo, en un arreglo que contenga 50'000,000 de elementos, este algoritmo realiza como máximo 26 comparaciones (en el peor de los casos).

El pseudocódigo de la versión recursiva de este algoritmo se muestra a continuación. En éste se



tienen como entradas un arreglo ordenado  $A$  y un valor buscado  $v$ , y se espera como salida un índice  $i$  tal que  $v = A[i]$  o nil.

BINARY-SEARCH ( $A, v, p, r$ )

**if**  $p \geq r$  and  $v \neq A[p]$  **then**

**return** nil

**end if**

$i \leftarrow A[\lfloor (r - p) / 2 \rfloor]$

**if**  $v = A[i]$  **then**

**return**  $i$

**else**

**if**  $v < A[i]$  **then**

**return** BINARY-SEARCH ( $A, v, p, i$ )

**else**

**return** BINARY-SEARCH ( $A, v, i, r$ )

**end if**

**end if**

# Capítulo 3

## Desarrollo del modelo de datos

Para el diseño y construcción de la bodega de datos y la implementación del modelo multidimensional se empleó el enfoque clásico top-down, construyendo un único repositorio de datos centralizado, y se siguió la arquitectura señalada por Vaisman and Zimányi (2014) para la implementación del modelo multidimensional, desarrollando en orden los siguientes componentes:

1. La capa oculta, que incluye la construcción de una base de datos intermedia, para llevar a cabo la extracción y transformación.
2. La capa de la bodega de datos, que se compone de la información transformada y cargada en la bodega de datos, utilizando las herramienta SQL Server Integration Services.
3. La capa OLAP, que consistió en el diseño e implementación del modelo multidimensional utilizando la herramienta OLAP de SQL Server Analysis Services.
4. La capa frontal, en la que se diseñaron diferentes consultas multidimensionales utilizando MDX, que serían explotadas posteriormente por un servicio web y un cliente móvil.

## 3.1. Extracción, transformación y carga

El proceso de extracción consistió en la selección de las fuentes de información para alimentar la bodega de datos. En el proceso de transformación, se realizaron diferentes operaciones que modificaron el contenido de algunos datos, así como operaciones de exclusión e integración, para que estos pudieran ser cargados en la bodega de datos. Finalmente, una vez que los datos fueron homologados e integrados, en el proceso de carga se llevó a cabo su almacenamiento en la bodega de datos.

### 3.1.1. Extracción de la información

Como fuente de datos primaria para el desarrollo del modelo de datos multidimensional se utilizó la base de datos de PBR de la GBIF. De esta forma, se pueden utilizar datos reales históricos para probar el hipercubo desarrollado.

Esta gran base de datos cuenta con una cantidad de registros de ocurrencias de especímenes superior a los 500 millones, para más de un millón 600 mil especies. De esta forma, aplicando una consulta relacional para extraer todos los registros de las ocurrencias junto con la información de las especies a la que pertenece cada ejemplar, daría como resultado un conjunto de datos con más de 500 millones de registros, lo cual se vuelve muy complicado manejar, dado el enorme tamaño de los archivos que tendrían que procesarse y almacenarse. Es por este motivo que se eligió como estrategia trabajar únicamente con un fragmento de esta base de datos, que cubriera geográficamente la región noroeste del territorio mexicano, aprovechando la herramienta para la exportación de datos de ocurrencias con que cuenta la página oficial de la GBIF. Esta herramienta, permite trazar un polígono sobre el mapa (ver fig. 3.1), de forma que se extraigan únicamente la información de las localidades de interés.

Utilizando esta herramienta, se delimitó el polígono con el conjunto de pares de coordenadas (longitud latitud):  $-117.773437, 32.694865$ ;  $-114.433593, 32.546813$ ;  $-108.984375, 29.840643$ ;  $-105.292968, 19.642587$ ;  $-113.730468, 20.797201$ ;  $-117.773437, 32.694865$ .

De esta forma, se creó un archivo de texto en formato CSV codificado en UTF-8 que contiene

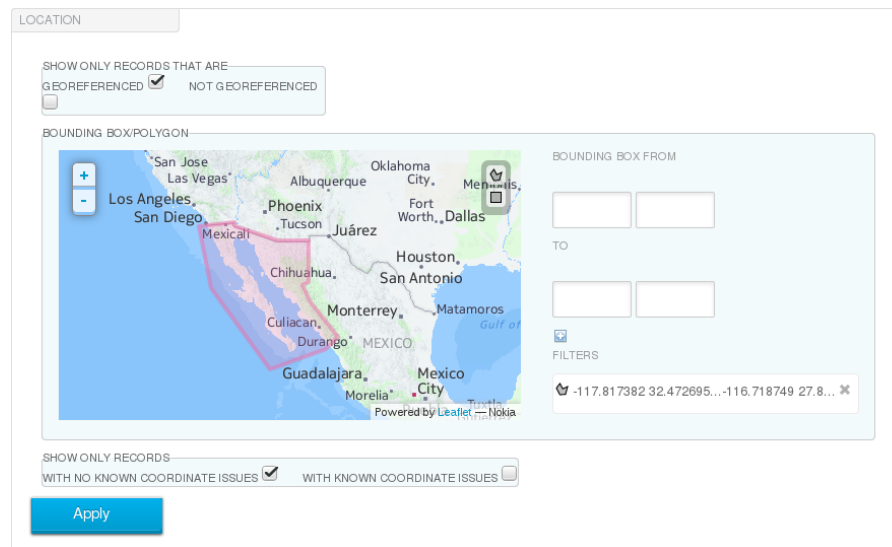


Figura 3.1: Captura de pantalla de la herramienta de exportación de datos del sitio oficial de la GBIF. En esta herramienta se pueden delimitar las localidades que serán consideradas para la exportación, de modo que se obtenga un conjunto de datos más reducido. Tomada de <http://www.gbif.org/occurrence/search>

los registros de las ocurrencias seleccionados. Así, se logró extraer un archivo que contiene un total de 804,462 registros, con un peso de 126.5 MB, lo cual permite manejar este conjunto de datos con mucha mayor facilidad, en cualquier SGBD moderno.

Entre los datos más relevantes que presenta cada registro contenido en este archivo encontramos:

**gbifID** Identificador único del registro de la ocurrencia del ejemplar

**decimalLatitude** Latitud decimal de la localización geográfica de la ocurrencia

**decimalLongitude** Longitud decimal de la localización geográfica de la ocurrencia

**locality** Nombre de la localidad en la que se encontró el ejemplar registrado. Este dato puede tratarse de una descripción geográfica del lugar, utilizando un punto de referencia conocido o simplemente el nombre de una población o de una localidad geográfica. Además de este dato de referencia geográfica, se incluyen otros relacionados a la ubicación de ejemplar, entre los que encontramos **county** (condado), **municipality** (municipio), **stateProvince** (estado o provincia), **countryCode** (código ISO2 del país en donde se encontró el ejemplar) y **continent** (nombre del continente donde se encontró el ejemplar).

**taxonKey** Llave taxonómica o identificador numérico del grupo de organismos al que pertenece el ejemplar. Por lo general, esta llave es el identificador de la especie a la que pertenece el ejemplar registrado. Sin embargo, debido a que la identificación de un organismo no siempre llega a nivel especie, sobre todo en los registros más antiguos, dadas las limitaciones a los que se enfrentaron los especialistas al digitalizarlos, esta llave puede también indicar el género, la familia, el orden o la clase del ejemplar. De modo que, a menor detalle en la información del registro, la clasificación será menos específica.

**species** Además de la llave de la especie, también se encuentra un campo que contiene el nombre de la especie a la que pertenece cada ejemplar, junto con otros campos para el resto de la clasificación taxonómica: **genus** (género), **family** (familia), **order** (orden), **class** (clase), **phylum** (filo) y **kingdom** (reino). En una base de datos relacional, siguiendo las reglas de normalización, todos estos campos deberían estar localizado en otras tablas, pero dado que el archivo es producto de una consulta relacional, cada registro cuenta con esta información, produciéndose de esta forma redundancias, que después serán solventadas con el diseño y construcción del cubo.

**eventDate** Fecha en la que ocurrió el evento de registro del ejemplar. Además, se cuenta con tres campos numéricos que desglosan esta fecha: **year**, **month** y **day**.

**basisOfRecord** Indica el origen del que se obtuvo el registro del ejemplar, pudiéndose tratar de un ejemplar capturado y preservado en una colección biológica, una observación registrada durante una investigación de campo, un ejemplar viviente perteneciente a un jardín botánico o zoológico, un registro fósil o un ejemplar registrado en la literatura científica.

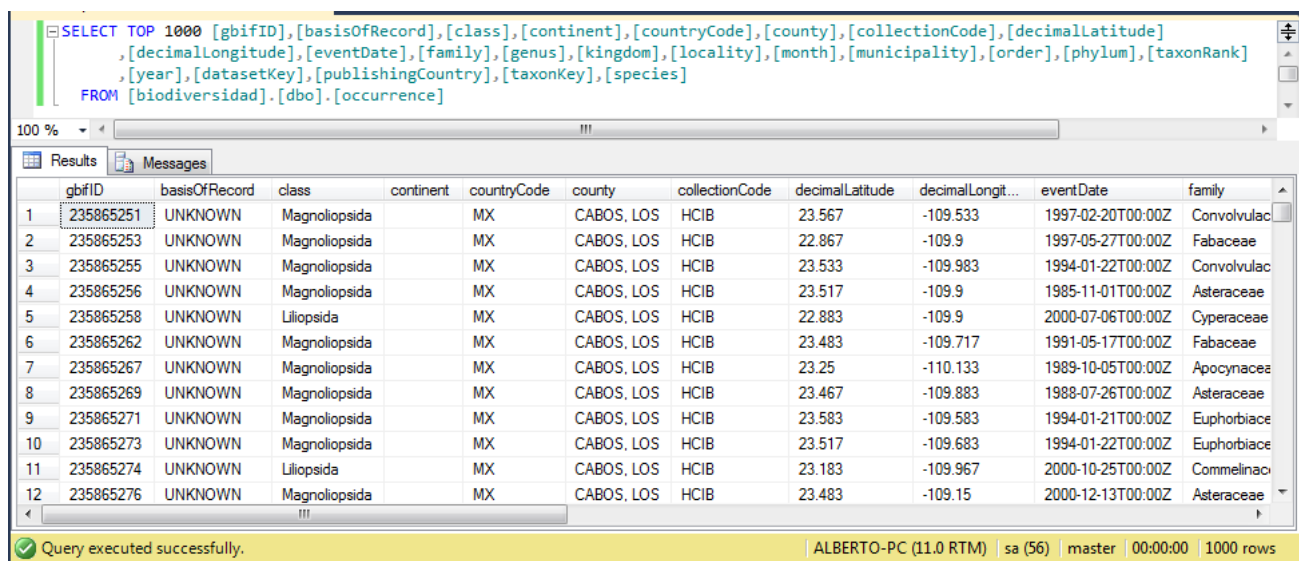
**datasetKey** Identificador alfanumérico de la colección biológica o base de datos de observaciones, en la que se encuentra el registro del ejemplar.

**collectionCode** Acrónimo de la colección en la que se encuentra el registro del ejemplar (en caso de tratarse de un espécimen preservado).

**publishingCountry** Código ISO2 del país al que pertenece la colección o institución que registró al ejemplar.

Finalmente, utilizando este archivo CSV se construyó una base de datos con el SGBD Microsoft SQL Server 2012, por medio su herramienta de exportación de datos para tablas. Gracias a este mecanismo se creó una primera tabla llamada `occurrence`, que contiene todos los PBR del conjunto de datos extraído.

En la fig. 3.2 se muestra una consulta a la tabla creada en la base de datos, visualizando únicamente los datos relevantes para el diseño de las dimensiones del cubo: identificador del registro, datos del grupo de organismos, datos geográficos, fecha del evento, quién recolectó el ejemplar y la base del registro.



```

SELECT TOP 1000 [gbifID],[basisOfRecord],[class],[continent],[countryCode],[county],[collectionCode],[decimalLatitude]
,[decimalLongitude],[eventDate],[family],[genus],[kingdom],[locality],[month],[municipality],[order],[phylum],[taxonRank]
,[year],[datasetKey],[publishingCountry],[taxonKey],[species]
FROM [biodiversidad].[dbo].[occurrence]

```

	gbifID	basisOfRecord	class	continent	countryCode	county	collectionCode	decimalLatitude	decimalLongit...	eventDate	family
1	235865251	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.567	-109.533	1997-02-20T00:00Z	Convolutac
2	235865253	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	22.867	-109.9	1997-05-27T00:00Z	Fabaceae
3	235865255	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.533	-109.983	1994-01-22T00:00Z	Convolutac
4	235865256	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.517	-109.9	1985-11-01T00:00Z	Asteraceae
5	235865258	UNKNOWN	Liliopsida		MX	CABOS, LOS	HCIB	22.883	-109.9	2000-07-06T00:00Z	Cyperaceae
6	235865262	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.483	-109.717	1991-05-17T00:00Z	Fabaceae
7	235865267	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.25	-110.133	1989-10-05T00:00Z	Apocynaceae
8	235865269	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.467	-109.883	1988-07-26T00:00Z	Asteraceae
9	235865271	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.583	-109.583	1994-01-21T00:00Z	Euphorbiaceae
10	235865273	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.517	-109.683	1994-01-22T00:00Z	Euphorbiaceae
11	235865274	UNKNOWN	Liliopsida		MX	CABOS, LOS	HCIB	23.183	-109.967	2000-10-25T00:00Z	Commelinac
12	235865276	UNKNOWN	Magnoliopsida		MX	CABOS, LOS	HCIB	23.483	-109.15	2000-12-13T00:00Z	Asteraceae

Query executed successfully. | ALBERTO-PC (11.0 RTM) | sa (56) | master | 00:00:00 | 1000 rows

Figura 3.2: Captura de una consulta realizada sobre la tabla `occurrence` creada para alimentar el hiper cubo diseñado en este proyecto. En esta visualización se pueden apreciar varios de los campos utilizados para definir las dimensiones del modelo multidimensional aplicado en la bodega de datos.

### 3.1.2. Transformación de la información

Al momento de consultar la tabla de ocurrencias de especímenes, encontramos que algunos de los campos de que dispone deben ser transformados para poder aprovecharse en el proceso de análisis al que serán incorporados dentro de la bodega de datos. Además, otros campos fueron excluidos por su poca relevancia para el modelo multidimensional, mientras que fue necesario

incluir nuevas tablas en la base de datos, que sustentarían algunas dimensiones del modelo. Todo esto previo a la carga de la información en la bodega de datos.

### 3.1.2.1. Limpieza de los datos

Encontramos que dentro de los datos geográficos, `stateProvince` presenta inconsistencias al momento de nombrar los estados de la república mexicana, apareciendo por ejemplo variantes del nombre de Baja California Sur, tales como `baja california sur`, `Baja California (sur)`, `BCS`, etc. (ver fig. 3.3). Esto debido a fallas de origen en la captura de los datos por parte de los responsables de su registro.

(No column name)	stateProvince	
6	7	Baha California Norte
7	20	baie de californie
8	21	Baja
9	3	Baja California
10	1	Baja California
11	3	Baja Calif
12	6	Baja calif suk
13	40	Baja calif sur
14	3	BAJA CALIF.
15	11	BAJA CALIF. NORTE
16	2	BAJA CALIF. SUR
17	161757	Baja California
18	30	Baja California (nor
19	1	Baja California (norte)
20	37	Baja California del Norte
21	44	Baja California del Sur

Query executed successfully. ALBERTO-PC (11.0 RTM) sa (51) biodiversidad 00:00:49 105 rows

Figura 3.3: Estado de los datos en el campo `stateProvince` dentro de la tabla de ocurrencias de ejemplares

Por lo tanto, fue necesario homologar este campo a través de una serie de comandos SQL para corregir estas inconsistencias. Como muestra, esta es la consulta para corregir los registros de Baja California Sur:

```
UPDATE occurrence
SET stateProvince = 'Baja California Sur'
WHERE
stateProvince COLLATE Latin1_General_CI_AI
LIKE '%baja %california %sur %'
OR
```

```
stateProvince COLLATE Latin1_General_CI_AI
LIKE '%baja %calif %su %'
OR
stateProvince COLLATE Latin1_General_CI_AI
LIKE '%bcs %';
```

Aplicando un procedimiento similar para los otros estados, quedaron corregidas las inconsistencias en el campo `stateProvince`.

### 3.1.2.2. Cálculo de la fecha de la ocurrencia

Por otro lado, el campo `eventDate`, consiste en una cadena que representa la fecha en formato ISO 8601 con el tiempo expresado en UTC, siendo este último no soportado por la función de transformación de fechas que incorpora SQL Server. Para facilitar el trabajo de la transformación, en lugar de usar este campo se utilizaron los tres campos numéricos que representan esta misma fecha: `year`, `month` y `day`, con los cuales fue posible crear un nuevo campo fecha utilizando una función del propio SQL Server.

Para lograr esto, primero se añadió el campo `occurrenceDate` de tipo `datetime` a la tabla, y después se ejecutó la instrucción:

```
SET DATEFORMAT ymd
UPDATE [dbo].[occurrence]
SET
[dbo].[occurrence].occurrenceDate = CONVERT(char(11),
CAST([dbo].[occurrence].[year] AS VARCHAR(4))
+ '/'
+ REPLICATE('0',
2 - LEN(CAST([dbo].[occurrence].[month] AS VARCHAR(2))))
+ CAST([dbo].[occurrence].[month] AS VARCHAR(2))
+ '/'
+ REPLICATE('0',
```



```

2 - LEN(CAST([dbo].[occurrence].[day] AS VARCHAR(2)))
+ CAST([dbo].[occurrence].[day] AS VARCHAR(2)), 111)
WHERE [dbo].[occurrence].[year] >= 1800

```

### 3.1.2.3. Exclusión de datos

Existieron varios campos que fueron excluidos de su carga en la base de datos, ya que eran de poca importancia para el diseño de las dimensiones del modelo, además de que en la mayoría de los registros se encontraban vacíos. Entre estos campos fueron los siguientes:

accessRights, bibliographicCitation, identifier, modified, references, rights, rightsHolder, source, acceptedNameUsage, associatedOccurrences, associatedReferences, associatedSequences, associatedTaxa, taxonRemarks, typeStatus, verbatimCoordinateSystem, verbatimDepth, verbatimElevation, verbatimEventDate, verbatimLocality, verbatimSRS, verbatimTaxonRank y waterBody.

Además, se eliminaron los registros que no contaban con la fecha de registro y que su clasificación no llegaba a nivel género, ya que no serían útiles al momento de implementar las dimensiones “Grupo de organismos” y “Tiempo” en el hiper cubo. Esto se logró con la instrucción:

```

DELETE [dbo].[occurrence]
WHERE
[dbo].[occurrence].[eventDate] = ''
OR [dbo].[occurrence].[genus] = ''

```

### 3.1.2.4. Incorporación de las regiones de interés y sus polígonos

Otra observación importante en los datos sobre localidades geográficas, es que encontramos registros que no están registrados para ningún estado de la República Mexicana. Para descartar que se tratara de un error, realizamos una consulta del campo `locality` en estos registros. El resultado, que se puede apreciar en la fig. 3.4, fue que la gran mayoría de las localidades se

encuentran en el mar, siendo éstas islas, zonas costeras o regiones mar adentro, ya sea del Golfo de California o del litoral del Pacífico Norte Mexicano.

```

SELECT [dbo].[occurrence].[locality], [dbo].[occurrence].[municipality]
FROM [dbo].[occurrence] WHERE [dbo].[occurrence].[stateProvince] = ''

```

	locality	municipality
66	Middle American Trench; 11.0 miles, 240° T from Cabo Alto; Isla Madre	
67	Gulf of California; 54.5 miles, 085° T from Pta. Arena	
68	Middle American Trench, 12 miles 242° T from Maria Cleofas Island Light	
69	2 mi. S of San Felipe, Gulf of California	
70	Gulf of California; 46.0 miles, 116° T from Isla Tortuga	
71	Gulf of California; Isla Camen	
72	18.0 Mi., 249 degrees T from Island Maria Cleofas Light	
73	Gulf of California; Ballenas Channel, 3 miles inside southern tip Angel de la...	
74	Gulf of California; 45.0 miles, 036° T from Pta. Arena Light	
75	Gulf of California; Piedras Gordas, Cerralbo Island (=Isla Cerralvo)	
76	Gulf of California; Isla Lobos	
77	12.3 miles, 207 degrees T from Cabo Alto, Maria Madre Isl.	
78	Gulf of California; 46.0 miles, 116° T from Isla Tortuga	
79	just S. Arenas Point in Gulf of California (ca. 1.5 mi.)	
80	Gulf of California; 54.5 miles, 085° T from Pta. Arena	

Query executed successfully. ALBERTO-PC (11.0 RTM) sa (53) biodiversidad 00:00:30 19141 rows

Figura 3.4: Resultado de la consulta de las localidades que no se encuentran dentro de un estado o municipio, se puede apreciar que muchas de ellas se encuentran dentro de regiones marinas.

Por otro lado, ninguno de los registros contó con información en su campo `municipality`, y existió un número importante de registros que tampoco contaban con la descripción de la localidad geográfica, apareciendo los campos `locality` y `county` vacíos. Además, en los registros que presentan una localidad geográfica, encontramos que muchos de los datos capturados en este mismo campo no poseen una estructura consistente: algunos nombres aparecen en inglés y otros en español, incluso en una misma localidad. Además, aún en el mismo idioma una misma localidad puede llamarse de distintas formas (p. e. Bahía de San Jorge - Isla Espiritu Santo, Bahía de San Gabriel - Isla Espiritu Santo). También se encontró que no todos los registros tienen el mismo nivel de detalle, ya que algunos utilizan puntos de referencia con unidades de distancia y dirección para describir la ubicación (p. e. Gulf of California; 26.7 miles 281°T from Island San Juanito Light), y otros sólo señalan el nombre de un lugar (p. e. Gulf of California o Isla Cerralvo).

Todos estos detalles dificultarían el aplicar una dimensión espacial en el modelo de datos empleando sólo estos campos, por lo que se recurrió a las coordenadas geográficas, presentes en todos los registros, para clasificar estos datos dentro de nuevas regiones.

Para lograrlo, fue necesario incorporar tres tablas adicionales a la base de datos: un catálogo de regiones, una tabla de los puntos (o vértices) de los polígonos de cada región, y una tabla

relación entre las regiones y la tabla de ocurrencias.

En la tabla de regiones (identificada como `region`) aparecen los nombres de las regiones geográficas de interés definidas en este proyecto, dentro de las cuales podemos clasificar una ocurrencia. La tabla de vértices (llamada `vertex`) está relacionada con la de regiones, de forma que contiene los puntos geográficos de los polígonos que conforman cada región. La tabla relación final (`occurrence_region`) es la que permite clasificar las ocurrencias dentro de las regiones. En la fig. 3.5 se puede visualizar a detalle la relación creada entre estas tablas.

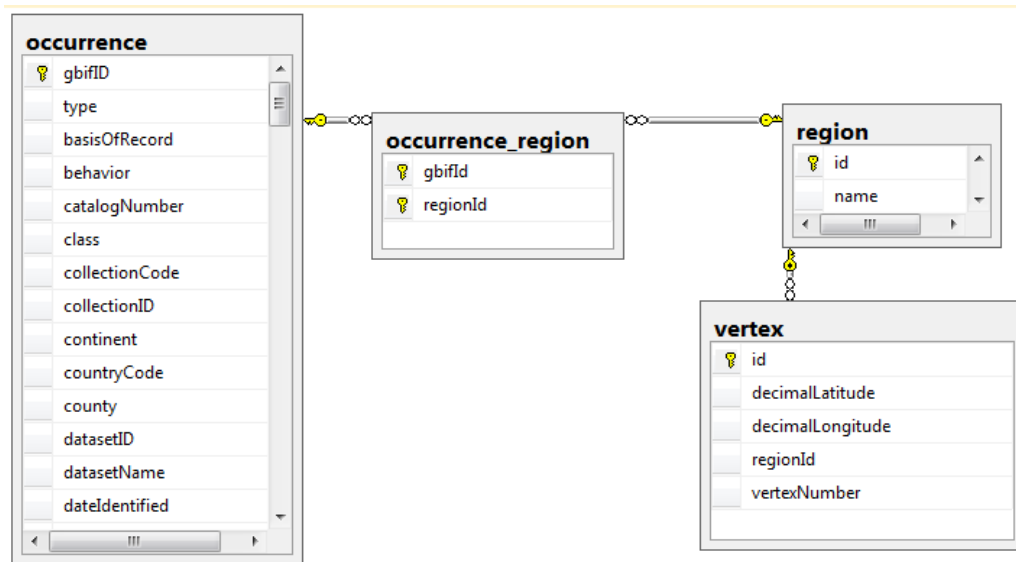


Figura 3.5: Diagrama Entidad Relación (ER) de la relación entre la tabla de ocurrencias y las nuevas tablas creadas para la representación de las regiones geográficas de interés.

El primer paso fue crear una región de interés para almacenarla en la base de datos. Para esto se empleó el software QGIS para definir el polígono de puntos geográficos que la conformarían (ver fig. 3.6). Los pares de coordenadas de esta región son -110.41902081 24.18063388; -110.32254662 24.18043833; -110.32129211 24.14860054; -110.41946034 24.14711427. Estos punto definen un cuadrilátero sobre la región conocida como “El Mogote” en la Bahía de La Paz.

Esta información se almacenó en la base de datos dentro de las tablas `region` y `vertex`, con los comandos:

```

INSERT INTO region (name) VALUES('El Mogote ');
INSERT INTO polygon
  
```

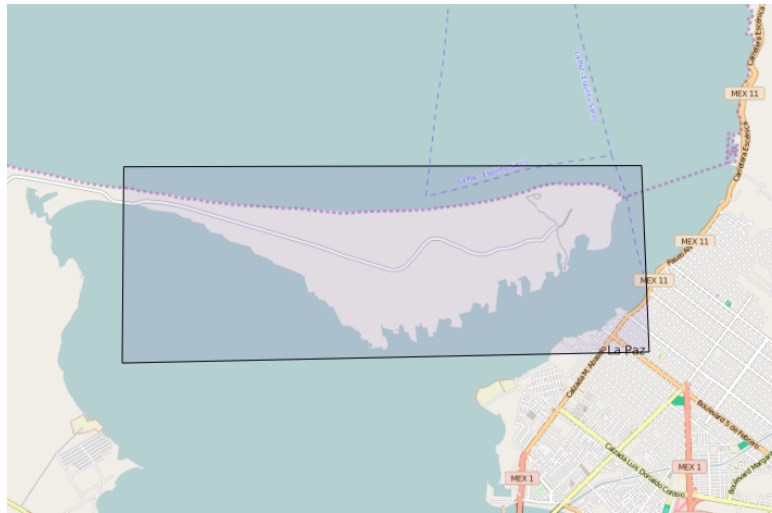


Figura 3.6: Visualización del polígono trazado utilizando QGIS, con el cual se creará la región de “El Mogote” dentro de la base de datos.

```
(decimalLongitude , decimalLatitude , regionId , vertexNumber)
```

```
VALUES
```

```
(-110.41902081, 24.18063388, 1, 1),
(-110.32254662, 24.18043833, 1, 2),
(-110.32129211, 24.14860054, 1, 3),
(-110.41946034, 24.14711427, 1, 4);
```

Finalmente, se programó el procedimiento almacenado para clasificar los puntos que cayeran dentro de esta región geográfica, utilizando las instrucciones:

```
CREATE PROCEDURE [dbo].[SearchOccurrencePoints]
@regionId as int
AS
BEGIN
    DECLARE @decimalLatitude real;
    DECLARE @decimalLongitude real;
    DECLARE @polygonString varchar(4096);
    DECLARE @lastPointString varchar(128);
    DECLARE @occurrenceCursor CURSOR;
    DECLARE @gbifId int;
```

```
DECLARE @found bit;
DECLARE @poly geography;
DECLARE @point geography;

SET NOCOUNT ON;

SELECT @polygonString = STUFF(
    (SELECT
        CONCAT( ', ',
            STR(decimalLongitude, 11, 6),
            ' ',
            STR(decimalLatitude, 11, 6))
    FROM vertex
    WHERE regionId = @regionId
    ORDER BY vertexNumber
    FOR XML PATH ( '' )
    , 1, 1, '' )
SELECT @lastPointString = CONCAT(
    ', ',
    STR(decimalLongitude, 11, 6),
    ' ',
    STR(decimalLatitude, 11, 6))
FROM vertex
WHERE regionId = @regionId
AND vertexNumber = 1;

SET @polygonString = CONCAT(@polygonString, @lastPointString);

SET @occurrenceCursor = CURSOR FOR
SELECT gbifID, decimalLatitude, decimalLongitude
FROM dbo.occurrence
```

```
OPEN @occurrenceCursor
FETCH NEXT FROM @occurrenceCursor
INTO @gbifId , @decimalLatitude , @decimalLongitude

WHILE @@FETCHSTATUS = 0
BEGIN
    SET @poly = geography::STGeomFromText(
        'POLYGON((' + @polygonString + '))' ,
        4326);
    SELECT
        @poly = (
            CASE
            WHEN @poly.EnvelopeAngle() > 90
            THEN @poly.ReorientObject()
            ELSE @poly END)
    SET @point = geography::Point(
        @decimalLatitude ,
        @decimalLongitude ,
        4326)
    SET @found = @poly.STIntersects(@point)
    IF @found = 1
    BEGIN
        INSERT INTO
            occurrence_region
            (gbifId , regionId)
            VALUES(@gbifId , @regionId)
    END
    FETCH NEXT FROM @occurrenceCursor
    INTO @gbifId , @decimalLatitude , @decimalLongitude
END;
CLOSE @occurrenceCursor;
```

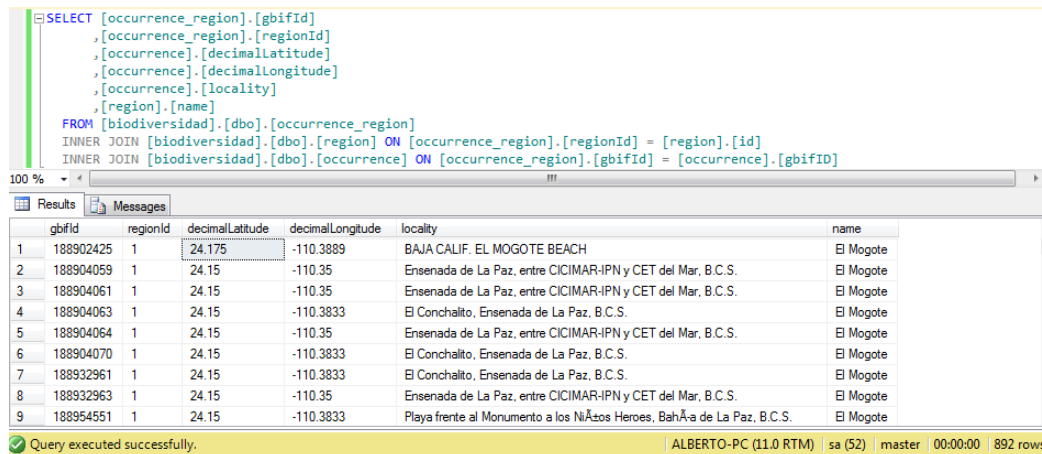
```
DEALLOCATE @occurrenceCursor ;
```

END

De esta forma se logran clasificar dentro de regiones las ocurrencias georreferenciadas invocando este procedimiento almacenado con la instrucción:

```
EXECUTE SearchOccurrencesPoints(1)
```

, donde el parámetro recibido es el identificador numérico de la región. Después de ejecutar la consulta, que tardó 8 minutos y 10 segundos en procesar los más de 600 mil registros de la tabla de ocurrencias, se lograron clasificar todos los puntos dentro del polígono definido para la región de “El Mogote”, tal y como se muestra en la fig. 3.7.



```
SELECT [occurrence_region].[gbifId]
, [occurrence_region].[regionId]
, [occurrence].[decimalLatitude]
, [occurrence].[decimalLongitude]
, [occurrence].[locality]
, [region].[name]
FROM [biodiversidad].[dbo].[occurrence_region]
INNER JOIN [biodiversidad].[dbo].[region] ON [occurrence_region].[regionId] = [region].[id]
INNER JOIN [biodiversidad].[dbo].[occurrence] ON [occurrence_region].[gbifId] = [occurrence].[gbifId]
```

gbifid	regionid	decimalLatitude	decimalLongitude	locality	name	
1	188902425	1	24.175	-110.3889	BAJA CALIF. EL MOGOTE BEACH	El Mogote
2	188904059	1	24.15	-110.35	Ensenada de La Paz, entre CICIMAR-IPN y CET del Mar, B.C.S.	El Mogote
3	188904061	1	24.15	-110.35	Ensenada de La Paz, entre CICIMAR-IPN y CET del Mar, B.C.S.	El Mogote
4	188904063	1	24.15	-110.3833	El Conchalito, Ensenada de La Paz, B.C.S.	El Mogote
5	188904064	1	24.15	-110.35	Ensenada de La Paz, entre CICIMAR-IPN y CET del Mar, B.C.S.	El Mogote
6	188904070	1	24.15	-110.3833	El Conchalito, Ensenada de La Paz, B.C.S.	El Mogote
7	188932961	1	24.15	-110.3833	El Conchalito, Ensenada de La Paz, B.C.S.	El Mogote
8	188932963	1	24.15	-110.35	Ensenada de La Paz, entre CICIMAR-IPN y CET del Mar, B.C.S.	El Mogote
9	188954551	1	24.15	-110.3833	Playa frente al Monumento a los Niños Heroes, Bahía de La Paz, B.C.S.	El Mogote

Query executed successfully. ALBERTO-PC (11.0 RTM) sa (52) | master | 00:00:00 | 892 rows

Figura 3.7: Captura de pantalla que muestra el resultado de la consulta de las ocurrencias clasificadas dentro de la región conocida como “El Mogote”. De todos los registros de la tabla de ocurrencias, se encontró que 862 pertenecen a esta región.

Se siguió este mismo procedimiento para crear los polígonos de las regiones de “Isla Espiritu Santo”, “Bahía Magdalena”, “Isla Cerralvo”, “Sierra de la Laguna” y “Bahía de La Paz”. Las instrucciones SQL para la creación del resto de las regiones se encuentran en el apéndice A de este documento.

### 3.1.2.5. Creación del catálogo de colecciones

Como se mencionó antes, los registros de las ocurrencias pueden hacer referencia a especímenes preservados dentro de una colección biológica o herbario; o bien pueden tratarse de registros de observaciones de especímenes vivos realizadas por una organización de investigadores. En cualquiera de los casos, estos registros contarán con un registro electrónico en una base de datos perteneciente al grupo que registró la ocurrencia; por lo que, para facilitar la descripción del proceso de transformación, se llamó de forma genérica a este conjunto de datos **colección**.

Dado que no se cuenta con el nombre completo de las colecciones en la tabla `occurrence`, fue necesario crear un catálogo que contuviera estos nombres, relacionándolo con cada registro de la tabla `occurrence` a través del campo `datasetKey`. Para lograrlo, se utilizó nuevamente la base de datos de la GBIF, sólo que esta vez la información extraída se obtuvo de un conjunto de archivos XML, correspondiendo cada uno a una colección particular. Para lograr alimentar la base de datos intermedia con esta información, se creó un programa en el lenguaje R, para extraer la información requerida y colocarla en un formato soportado por el SGBD SQL Server. Se escogió el lenguaje R como herramienta por la gran facilidad que ofrece para leer grandes conjuntos de datos y procesar la información que contienen. De esta forma, se extrajeron de los archivos XML las claves y nombres las colecciones, creando un archivo CSV que, una vez transformado en una hoja de cálculo Excel, pudo ser importado dentro de la base de datos.

La tabla creada dentro de SQL Server, llamada `dataset`, contuvo únicamente tres campos:

**datasetKey** Identificador alfanumérico de la colección.

**datasetName** Nombre de la colección.

**publishingCountryId** Identificador numérico del país de la institución responsable de la publicación del registro de la colección, el cual hace referencia a la tabla `country` que se detalla en la sección 3.1.2.6.

El resultado final se puede apreciar en la fig. 3.8.



select dataset.sql ...iodiversidad (sa (53)) X

```

SELECT TOP 1000 [datasetKey]
, [datasetName]
FROM [biodiversidad].[dbo].[dataset]

```

100 %

	datasetKey	datasetName
1	005eb8d8-ed94-41be-89cf-e3115a9058e4	Field Museum of Natural History (Zoology) Invertebrate Collection
2	0096dfc0-9925-47ef-9700-9b77814295f1	Bioimages
3	0214a6a7-898f-4ee8-b888-0be60ecde81f	Molluscs collection (IM) of the Muséum national d'Histoire naturelle (MNHN - Paris)
4	02242d2f-b43f-44d1-9e53-4c51af461f5b	Oregon State University Herpetological Collection
5	0348540a-e644-4496-89d3-c257da9ad776	Marie-Victorin Herbarium (MT)
6	07d0d79-4883-435f-bba1-58fef110cd13	University of British Columbia Herbarium (UBC) - Vascular Plant Collection
7	0943f690-fde5-11dd-83f4-b8a03c50a862	Phanerogamic Botanical Collections (S)
8	0991f9af-8ff9-4e34-a8b9-990bb522dc0c	Mollusca collection of National Museum of Nature and Science
9	09c4287e-e6d5-4552-a07b-ff8a00833d8	MVZ Herp Collection (Arctos)
10	0e14d118-2af4-4e6f-8e32-6446620e12d4	Zoological Museum Amsterdam, University of Amsterdam (AM), Mammals

Query executed successfully. ALBERTO-PC (11.0 RTM) sa (53) biodiversidad 00:00:00 448 rows

Figura 3.8: Consulta de la tabla dataset que contiene los nombres de las colecciones que guardan los ejemplares registrados en la tabla occurrence.

### 3.1.2.6. Inclusión de los catálogos de estados y países

Para poder analizar la información del lugar de las ocurrencias, en los registros con información del país y del estado o provincia donde se realizó la captura u observación del ejemplar, se añadieron dos catálogos adicionales. Como primer paso, se creó una tabla de nombres de países con sus claves ISO, ya que los campos `countryCode` y `publishingCountry` contienen únicamente la clave ISO2 del país donde se localizó al ejemplar y del país al que pertenece la organización que lo hace público, respectivamente. Dado que estos datos son relevantes para la definición de las dimensiones del modelo, resulta mucho más útil para el usuario poder consultar estos países con sus nombres completos. Junto al catálogo anterior, como se segundo paso se incorporó una tabla que incluyera un catálogo de estados/provincias, que serviría para identificar el estado o provincia donde se tuvo la ocurrencia de los ejemplares, que a su vez se relacionaría con la tabla de países, para ubicar cada estado o provincia dentro de un país.

De esta forma, se creó la tabla `country`, que cuenta con los siguientes campos:

**id** Identificador numérico del país.

**iso2** Clave ISO2 que identifica el país (de dos caracteres alfabéticos).

**iso3** Clave ISO3 que identifica el país (de tres caracteres alfabéticos).

**shortName** Nombre corto del país, por ejemplo México, Estados Unidos, etc.

**longName** Nombre largo del país, por ejemplo Estados Unidos Mexicanos, United States of America, etc.

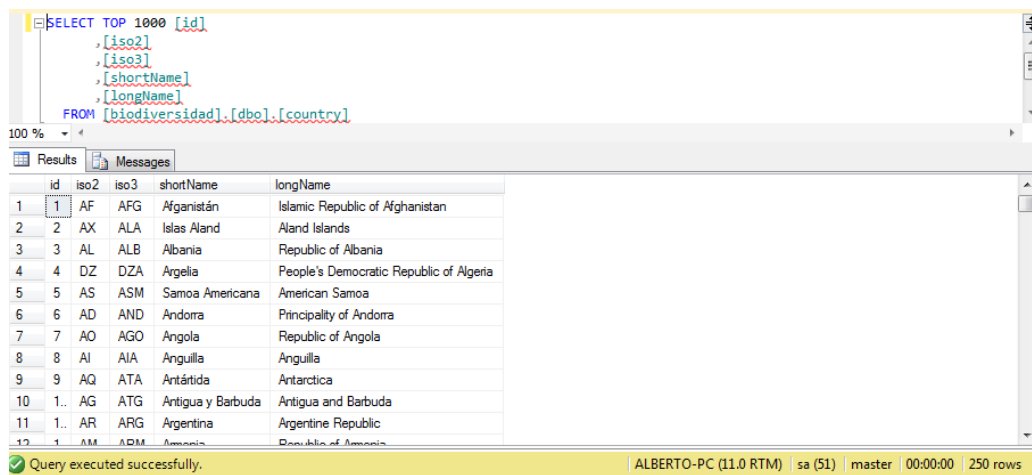
Así como la tabla `state_province`, que cuenta con los campos:

**id** Identificador numérico del estado/provincia.

**name** Nombre del estado o provincia, obtenido de la tabla `occurrence`.

**countryId** Identificador numérico del país donde se dio la ocurrencia del ejemplar, que hace referencia a la tabla `country`.

Para completar su información, se ejecutó un script SQL para insertar los datos de los países y estados. El resultado final se puede observar en la fig. 3.9.



The screenshot shows a SQL query window with the following query:

```
SELECT TOP 1000 [id]
, [iso2]
, [iso3]
, [shortName]
, [longName]
FROM [biodiversidad].[dbo].[country]
```

The results window displays a table with the following columns: id, iso2, iso3, shortName, and longName. The data is as follows:

id	iso2	iso3	shortName	longName
1	AF	AFG	Afganistán	Islamic Republic of Afghanistan
2	AX	ALA	Islas Aland	Aland Islands
3	3	ALB	Albania	Republic of Albania
4	4	DZA	Argelia	People's Democratic Republic of Algeria
5	5	ASM	Samoa Americana	American Samoa
6	6	AND	Andorra	Principality of Andorra
7	7	AGO	Angola	Republic of Angola
8	8	AIA	Anguilla	Anguilla
9	9	ATA	Antártida	Antarctica
10	1..	ATG	Antigua y Barbuda	Antigua and Barbuda
11	1..	ARG	Argentina	Argentine Republic
12	1	ARM	Armenia	Republic of Armenia

The status bar at the bottom indicates: Query executed successfully. ALBERTO-PC (11.0 RTM) | sa (51) | master | 00:00:00 | 250 rows

Figura 3.9: Consulta del catálogo de países creado para relacionarse con las tablas `dataset` y `state_province`.

De esta manera, la base de datos quedó modelada como se muestra en la fig. 3.10.

### 3.1.3. Carga de la información

Una vez terminada la transformación de los datos, el siguiente paso consistió en llevar a cabo el almacenamiento de la información en la bodega de datos, para así poder construir un hipercubo

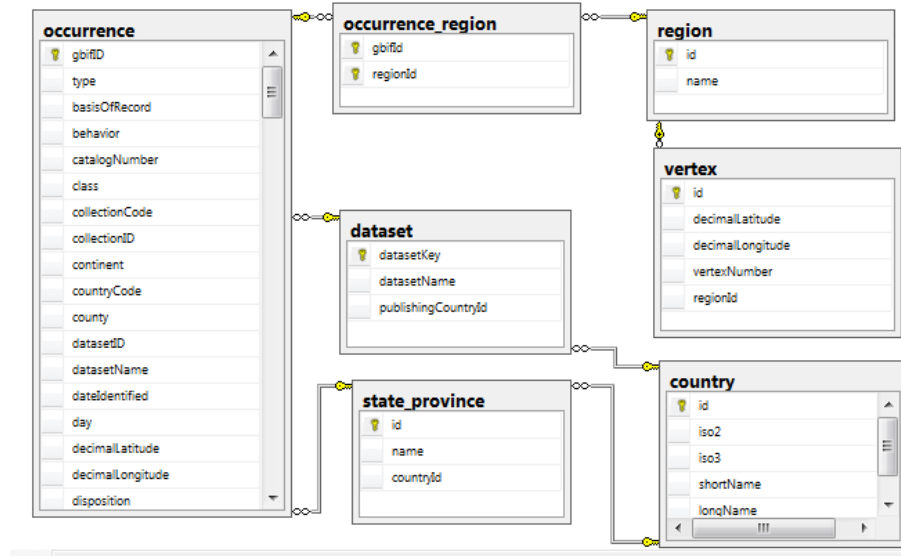


Figura 3.10: Diagrama entidad-relación del diseño de la base de datos intermedia construida para la creación de la bodega de datos. En él se puede apreciar la inclusión de los catálogos de colecciones, estados y países.

que implementara el modelo multidimensional diseñado.

Para efectuar esta tarea se utilizó la herramienta Microsoft SQL Server Integration Services, a través de la aplicación Visual Studio 2010 Shell. La bodega de datos se creó a partir de un nuevo proyecto dentro de este entorno. Para la carga de la información se utilizó una conexión al servidor de base de datos SQL Server, por medio del SQL Native Client 11.1 (SQLNCLI11.1). Una vez conectada a la fuente de datos (DataSource) se creó una vista de la fuente de datos (Data Source View), en la que se seleccionaron las tablas que serían incorporadas a la bodega de datos. Éstas fueron: `occurrence`, `region`, `occurrence_region`, `vertex`, `state_province`, `country` y `dataset`. El resultado final de la carga se puede apreciar en la fig. 3.11.

## 3.2. Diseño del modelo multidimensional

Una vez completada la capa de la bodega de datos, se trabajó en la capa OLAP, diseñando el modelo de datos multidimensional a través de la herramienta OLAP de Microsoft SQL Analysis Services.

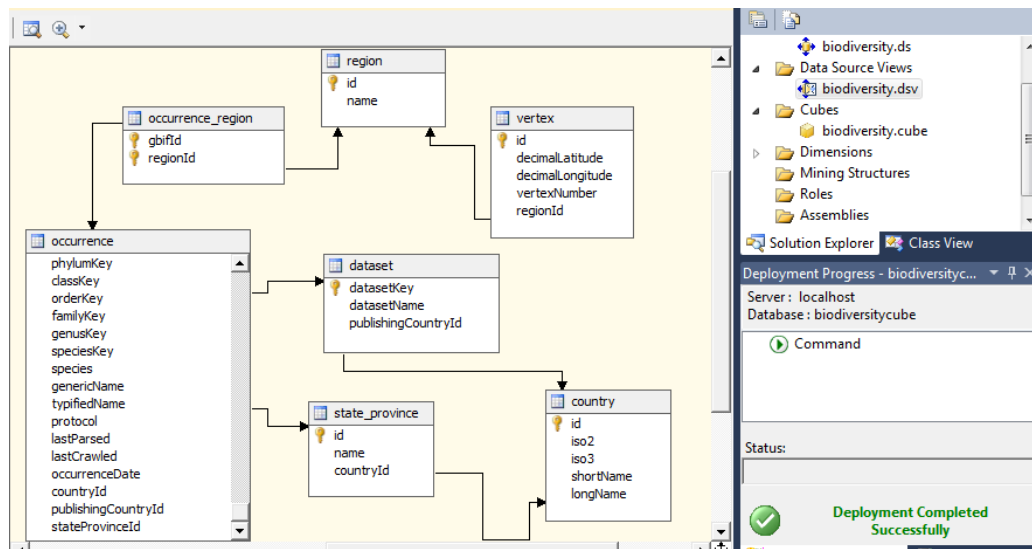


Figura 3.11: Visualización de las tablas cargadas dentro de la bodega de datos creada con la herramienta Microsoft SQL Server Integration Services

### 3.2.1. La tabla de hechos

La parte central del diseño multidimensional es la definición la tabla de hechos, ya que dentro de ésta encontramos el objeto del análisis (Vaisman and Zimányi, 2014). Para este proyecto, el punto medular es la información acerca de las ocurrencias de organismos, por lo que la tabla de hechos se obtuvo a partir de la tabla *occurrence*, que se puede visualizar en la fig. 3.12. Cada registro de esta tabla corresponde a la ocurrencia de un único ejemplar, registrándose cada uno como un evento único realizado por un grupo de investigadores, en una fecha específica y en un punto geográfico particular. Es posible, además, que varios registros correspondan a un mismo evento de investigación que obtuvo varios ejemplares en el mismo sitio, de forma que los datos contenidos en los campos *eventDate*, *datasetKey*, *decimalLatitude* y *decimalLongitude* serán los mismos para estos registros.

De esta forma, un hecho que podríamos extraer de esta tabla sería:

Se encontraron cinco organismos de la especie *Scarus ghobban* en la localidad de El Corralito, Isla Espíritu Santo, el día 15 de mayo de 2005 que fueron recolectados por los investigadores de la Colección Ictiológica del CICIMAR-IPN.

Si reemplazamos los datos de este hecho por variables, obtendríamos el siguiente enunciado:

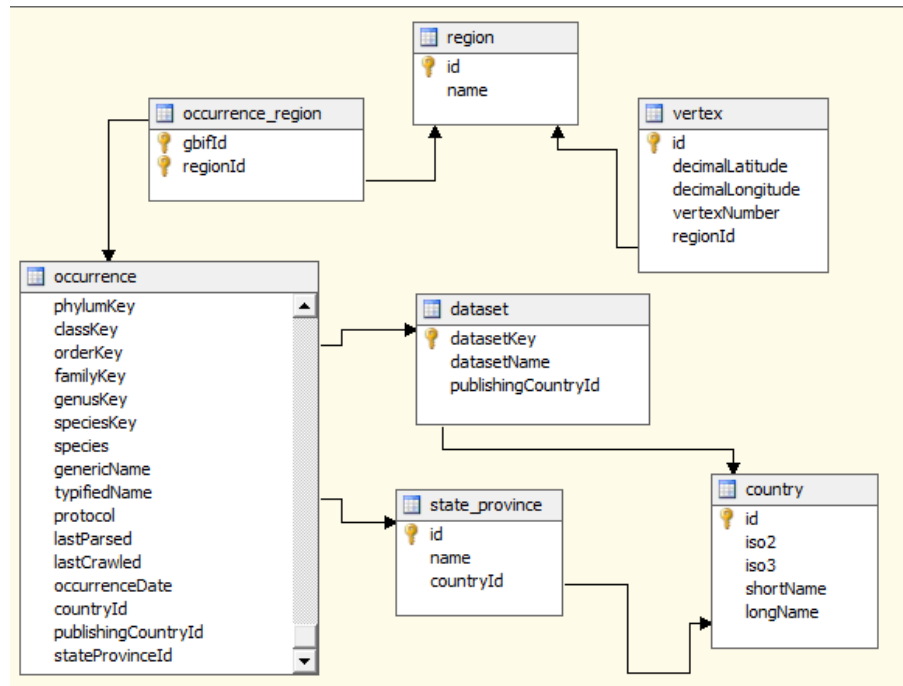


Figura 3.12: Diagrama entidad-relación de las tablas cargadas en la bodega de datos, donde se muestra la tabla `occurrence` que representará la tabla de hechos en este modelo de datos.

Se encontraron  $n$  organismos de la especie  $e$  en la localidad de  $l$ , el día  $d$ , que fueron recolectados por los investigadores de  $c$ .

Entonces, la variable  $n$  corresponde a una métrica, es decir, lo que estamos midiendo o cuantificando, mientras que las variables  $e$ ,  $l$ ,  $d$  y  $c$ , corresponderían a las características particulares de cada hecho, a partir de las cuales podemos definir las dimensiones de los hechos.

### 3.2.2. Definición de métricas

La métrica fundamental para este diseño es el número de ejemplares registrados en los eventos de investigación. Esta métrica se calcula directamente a partir del número de registros contabilizados en la tabla `occurrence`. Por ejemplo, la métrica para la especie *Scarus ghobban*, sería igual al número de registros que hacen referencia a esta especie.

### 3.2.3. Diseño lógico del modelo de datos

El tipo de esquema utilizado para representar la tabla de hechos y las dimensiones de este modelo es el esquema en copo de nieve, ya que se cuenta con una tabla de hechos central relacionada con las dimensiones diseñadas, en las cuales encontramos tanto tablas normalizadas como otras con redundancias (Vaisman and Zimányi, 2014).

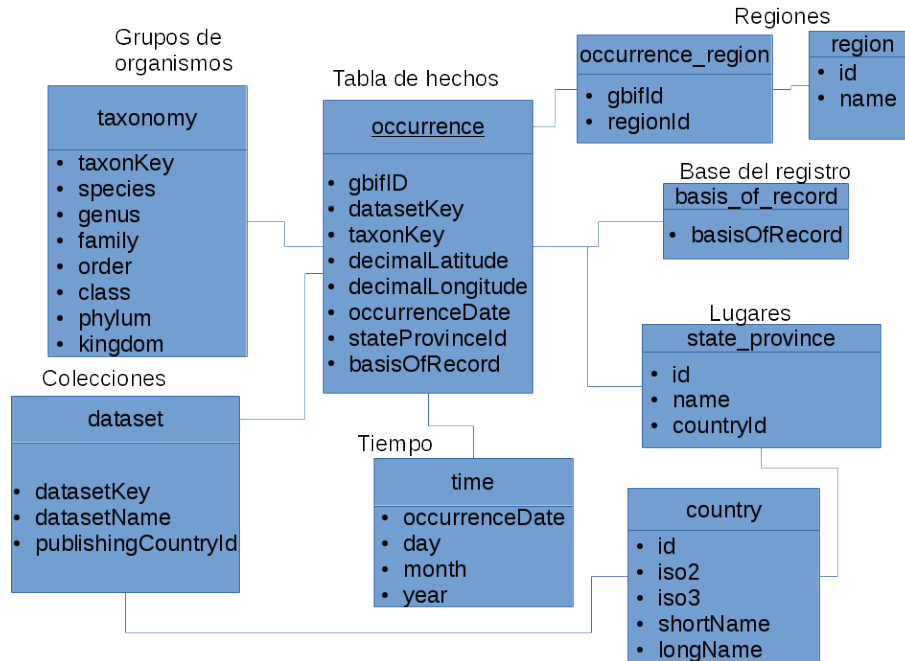


Figura 3.13: Esquema del modelo multidimensional diseñado, en el que se aprecian las seis grandes dimensiones que describen la información: grupo de organismos, lugares, regiones, colecciones, base del registro y tiempo.

En la fig. 3.13 se muestran las dimensiones detectadas siguiendo este tipo de diseño, de forma que observamos diferencias con el diagrama entidad-relación de la fig. 3.12. Estas diferencias se deben a que algunas de las dimensiones se obtendrán a partir de datos contenidos en la tabla *occurrence*, la cual no se encuentra totalmente normalizada, por lo cual encontramos que, por citar un ejemplo, los datos correspondientes a la dimensión de grupos de organismos (tabla *taxonomy*), en realidad se encuentran contenidos en los campos *taxonKey*, *species*, *genus*, *family*, *order*, *class*, *phylum*, y *kingdom*, todo esto dentro de la misma tabla *occurrence*, y se enlazan con la tabla de hechos a través de la llave *taxonKey*. Por otro lado, observamos también que la mayoría de las tablas de dimensiones están normalizadas, exceptuando la dimensión del

grupo de organismos, estando contenidos dentro de la misma todos los atributos de la jerarquía que componen a esta dimensión, debido a que esta es la forma en la cual encontramos estos datos en la tabla `occurrence` original.

### 3.2.4. Definición de las dimensiones y sus jerarquías

Para este modelo de datos se crearon seis dimensiones para el análisis de la información:

**Grupos de organismos** En esta dimensión, se analizan los datos a partir del grupo de organismos al que pertenece cada ejemplar registrado. Todo espécimen pertenecerá a una especie o género particular (algunos registros no cuentan con la especie), y esta clasificación pertenecerá a una jerarquía taxonómica ascendente, es decir: la especie pertenece a un género, el género a una familia, la familia a un orden, el orden a una clase, la clase a un filo y el filo a un reino. Esta dimensión se llamó “Taxonomy” dentro de la bodega de datos, construyéndose únicamente a partir de los datos de la tabla `occurrence`. Los atributos que la conforman son:

**Kingdom** Reino de la especie.

**Phylum** Filo de la especie.

**Class** Clase de la especie.

**Order** Orden de la especie.

**Family** Familia de la especie.

**Genus** Género de la especie.

**Species** Nombre de la especie.

En la fig. 3.14 se puede apreciar la jerarquía existente dentro de esta dimensión.

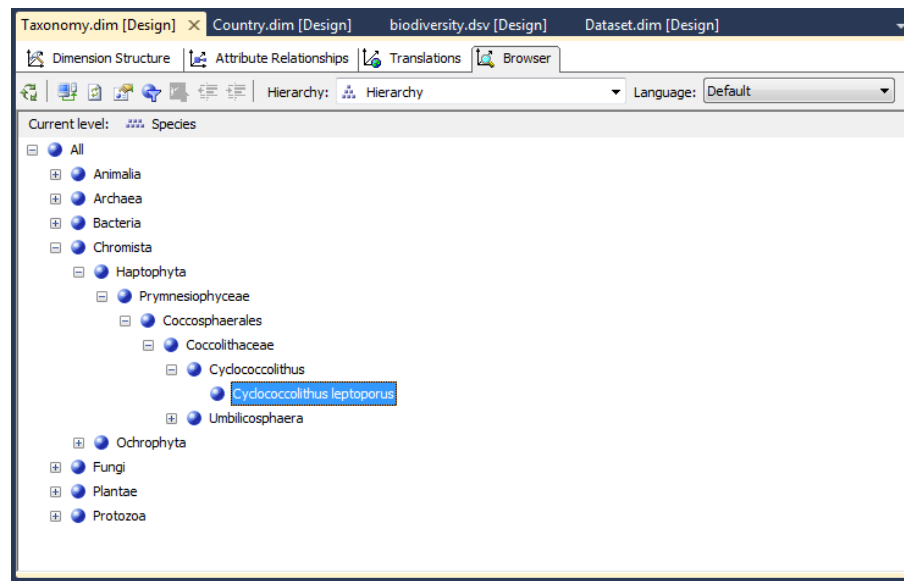


Figura 3.14: Visualización de la dimensión “Taxonomy”, que fue creada para clasificar los grupos de organismos. En este gráfico se aprecia la jerarquía que poseen los organismos del reino Chromista, yendo desde esta clasificación general hasta las clasificaciones particulares por especie.

**Lugares** Esta dimensión analiza los datos a partir de las localidades geográficas a las que hacen referencia los registros en la tabla `occurrence`. Estos datos forman una jerarquía que va desde el estado/provincia hasta el país de la ocurrencia. No se incluyó el campo `locality` en esta jerarquía, debido a la gran cantidad de inconsistencias que impiden utilizarlo para agrupar las ocurrencias de una misma localidad. Esta dimensión se llamó “Place” dentro de la bodega de datos, construyéndose a partir de los datos de la tabla `occurrence` y su relación con la tabla `state_province` a partir la llave foránea `stateProvinceId`. Los atributos que conforman esta dimensión son:

**Country Name** Nombre corto del país de la ocurrencia.

**State Province** Nombre del estado o provincia de la ocurrencia.

En la fig. 3.15 se puede apreciar la jerarquía que contiene.



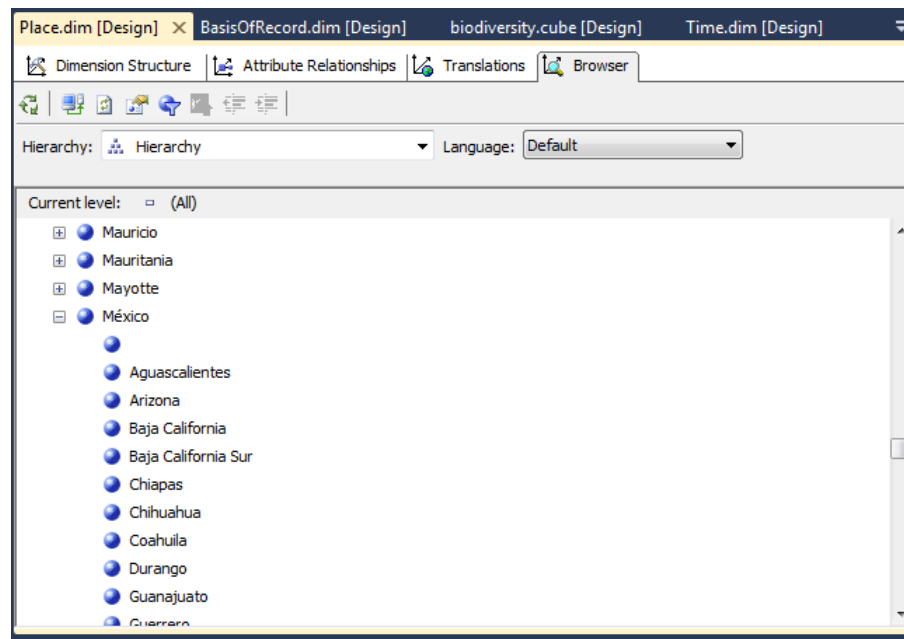


Figura 3.15: Visualización de la dimensión “Place”, que fue creada para analizar los lugares de las ocurrencias. En este gráfico se aprecia la jerarquía de los estados pertenecientes a la República Mexicana.

**Colecciones** Esta dimensión permite analizar la información de acuerdo al centro de investigación u organización, responsable de registrar la ocurrencia de un ejemplar. Esta dimensión se llamó “Dataset” dentro de la bodega de datos, utilizando las tabla `occurrence` y `dataset` para su construcción y relacionando ambas tablas a partir de la llave foránea `datasetKey`. Esta dimensión cuenta con una jerarquía de sólo dos niveles, con los atributos:

**Country Name** Nombre corto del país al que pertenece la institución u organización que posee la colección.

**Dataset Name** Nombre de la colección que posee el registro de la ocurrencia.

En la fig. 3.16 se puede apreciar la jerarquía de los atributos que posee.

**Regiones** Las regiones permiten analizar la información de las ocurrencias dentro de las áreas geográficas definidas por el usuario, a partir de los polígonos de puntos geográficos creados durante el proceso de transformación de los datos. Debido a que esta dimensión se obtuvo a

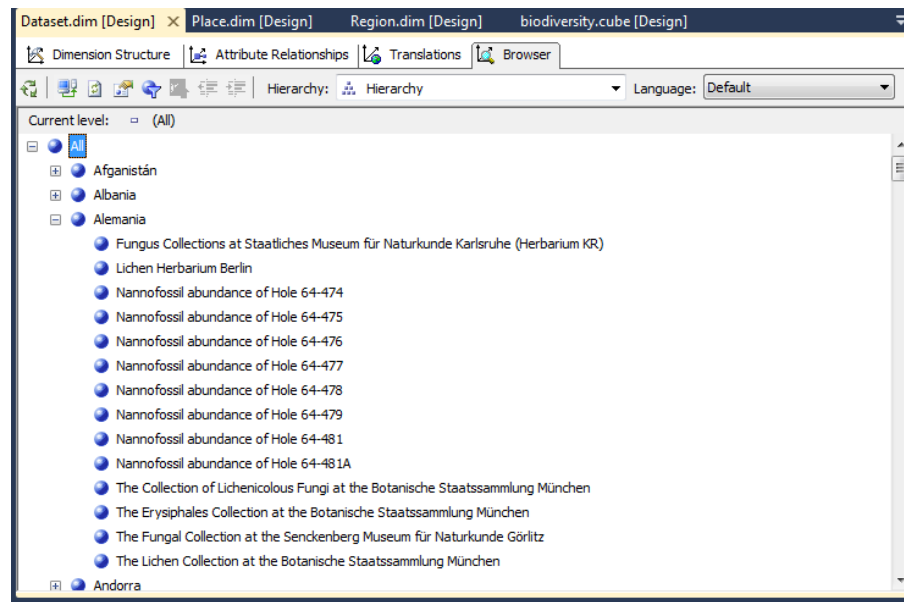


Figura 3.16: Visualización de la dimensión “Dataset”, creada para analizar las colecciones biológicas o proyectos de investigación que registraron las ocurrencias. En este gráfico se aprecian las colecciones que posee Alemania.

partir de la relación entre las tablas `occurrence` y `region`, que es del tipo *muchos a muchos* (a diferencia de las relaciones anteriores, que eran *uno a muchos*), fue necesario definir un grupo de métricas intermedio a partir de la tabla relación `occurrence_region` para, de esta forma, poder enlazar las regiones con las ocurrencias de ejemplares. Este grupo intermedio de métricas se llamó “Occurrence Region” dentro de la bodega de datos. Una vez incluido, se pudo crear la dimensión llamada “Region” a partir de la tabla `region`. Finalmente, esta dimensión se enlazó con la tabla de hechos, a través de una relación de dimensiones *muchos a muchos*. Esta dimensión posee únicamente un atributo y ninguna jerarquía:

**Region Name** Nombre de la región geográfica.

Los valores del atributo de esta dimensión se muestran en la fig. 3.17.

**Tiempo** Esta dimensión permite analizar la información a partir de sus datos temporales, es decir, el día, mes y año de cada ocurrencia. Para crear esta dimensión, se utilizó la función especial para la inclusión de una dimensión temporal que proporciona SQL Server Analysis

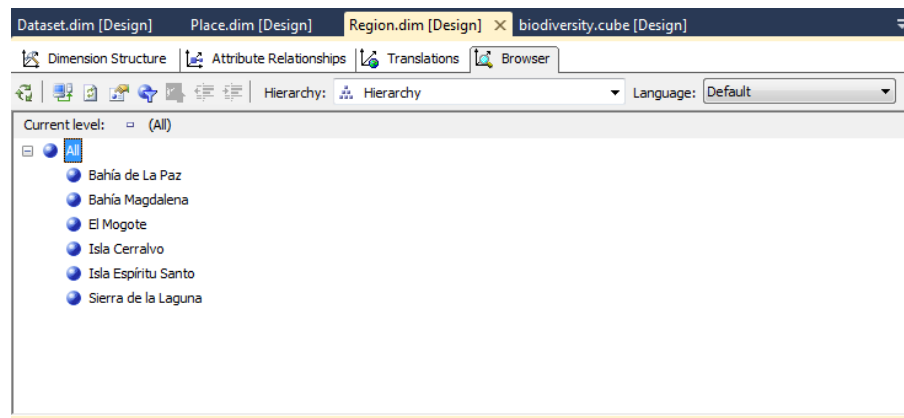


Figura 3.17: Visualización de los valores que conforman la dimensión “Region” en donde sólo se tiene un número pequeño de regiones definidas por el usuario, debido a lo tardado que resulta clasificar las ocurrencias dentro de cada nueva región.

Services. Esta función creó una tabla que posee la información de todas las fechas ocurridas dentro de un rango de tiempo definido, incluyendo datos del calendario para cada fecha (como el nombre del mes con su año, el semestre o el día de la semana). Además, durante su diseño se definió una jerarquía de tiempo para los niveles año, mes y día. De esta forma se mejora el desempeño de las consultas MDX, al no tener que realizar operaciones adicionales para visualizar las métricas, en los distintos niveles de detalle que posee esta dimensión. Como resultado final se incorporó una nueva tabla al diseño llamada `Time`, que se relacionó con la tabla `occurrence` a partir del campo `occurrenceDate`, como se aprecia en la fig. 3.18. Esta dimensión temporal se llamó “Time” dentro de la bodega de datos. En la fig. 3.19 se puede apreciar la jerarquía creada.

Debido a que las bodegas de datos almacenan información histórica, la dimensión del tiempo está presente en casi todos los diseños de bodegas de datos (Vaisman and Zimányi, 2014). En este modelo en particular, se incluyó esta dimensión debido al gran valor que brinda el poseer una perspectiva del cambio de la información a lo largo del tiempo.

**Base del registro** Esta última dimensión se utiliza para analizar la información, de acuerdo a la forma en la que se obtuvo de registro del organismo: ya sea por observación, a partir de un ejemplar vivo, por la preservación de un espécimen, a partir del registro fósil o basándose en la

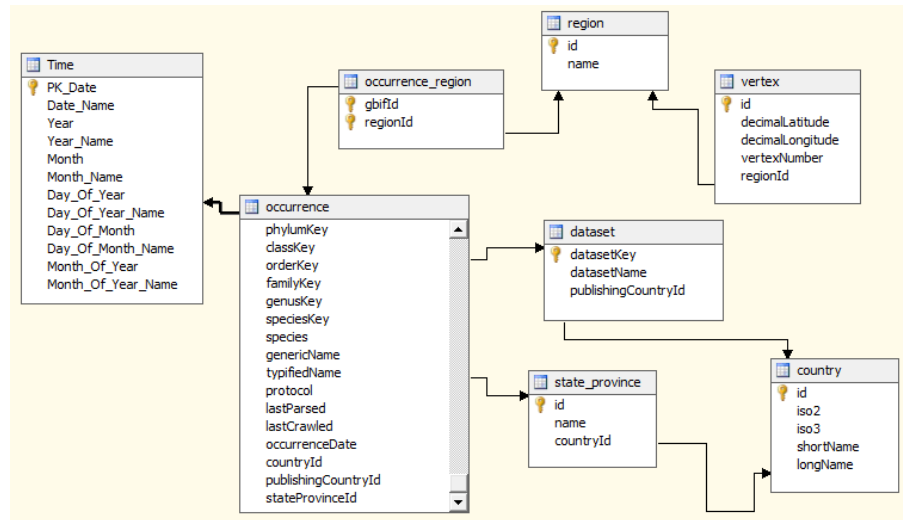


Figura 3.18: Visualización de la tabla de tiempo llamada time después de ser cargada dentro de la bodega de datos.

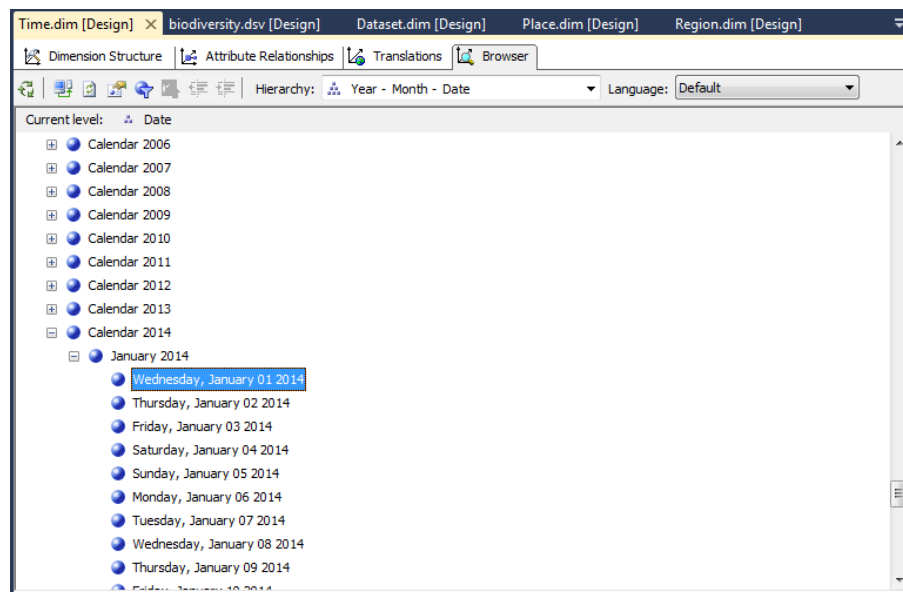


Figura 3.19: Visualización de la jerarquía de los datos que conforman la dimensión “Tiempo”, en donde se muestran los días que conforman el mes de enero del año 2014.

literatura. Esta dimensión se llamo “BasisOfRecord” y se compone de un único atributo:

**Basis Of Record** Base del registro.

Los datos que conforman ésta dimensión se muestran en la fig. 3.17.

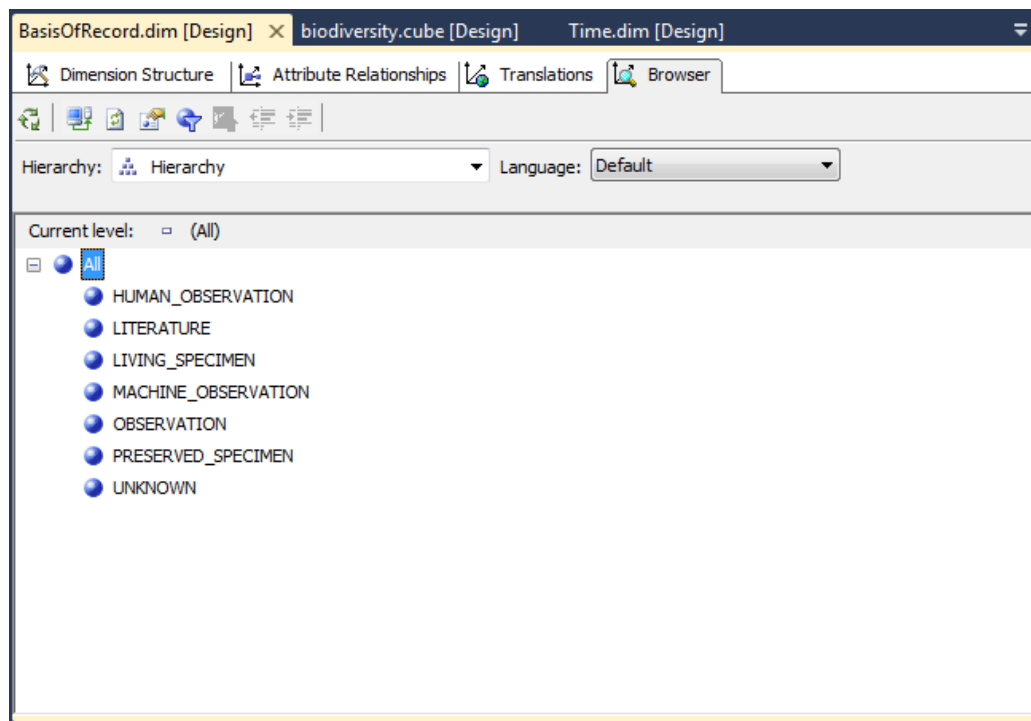


Figura 3.20: Visualización de los datos que componen la dimensión “BasisOfRecord”

### 3.2.5. Prueba de las dimensiones

Finalmente, para probar que las dimensiones fueron creadas correctamente dentro del hipercubo, utilizamos la herramienta de exploración para cubos de datos que incluye Visual Studio 2010 Shell. Por medio de la interfaz gráfica de esta herramienta es posible realizar consultas sencillas sobre las métricas del modelo multidimensional, combinando la diferentes dimensiones definidas. En la fig. 3.21 se observa una consulta sobre la métrica de número de ocurrencias, combinando tres dimensiones: la región, el mes del año (tiempo) y el reino (grupos de organismos).

Region Name	Kingdom	Month	Occurrence Count
Bahía de La Paz	Animalia	April 1880	1
Bahía de La Paz	Animalia	March 1882	1
Bahía de La Paz	Animalia	May 1882	2
Bahía de La Paz	Animalia	April 1888	81
Bahía de La Paz	Animalia	March 1889	13
Bahía de La Paz	Animalia	April 1889	4
Bahía de La Paz	Animalia	December 1894	1
Bahía de La Paz	Animalia	February 1895	1
Bahía de La Paz	Animalia	January 1905	3
Bahía de La Paz	Animalia	March 1911	1
Bahía de La Paz	Animalia	June 1919	1
Bahía de La Paz	Animalia	September 1919	1

Figura 3.21: Captura de la visualización de datos dentro del hipercubo construido. Aquí aparece una tabla que muestra el cálculo del número de organismos capturados (métrica del cubo) para las tres dimensiones de la información: la región, el mes del año (tiempo) y el reino (grupos de organismos).

### 3.3. Diseño de las consultas multidimensionales

Mediante las consultas multidimensionales podemos realizar cálculos sobre la información contenida en la bodega de datos, y visualizarla desde diferentes perspectivas. Los cálculos estarán relacionados con la métrica definida, mientras que las perspectivas dependerán de las dimensiones incluidas en el modelo, pudiendo hacer diferentes combinaciones entre las dimensiones para obtener nuevas perspectivas de la información, y moverse además entre las jerarquías construidas para obtener diferentes niveles de detalle.

Para construir estas consultas empleamos el lenguaje MDX implementado por el sistema OLAP de SQL Server Analysis Services. Una vez creadas las consultas, se hizo uso de ellas para el desarrollo de la aplicación móvil final.

Para facilitar la comprensión de las consultas multidimensionales, podemos visualizar gráficamente las dimensiones del modelo dentro de un cubo multidimensional, como el que se muestra en la fig. 3.22. Las distintas dimensiones aparecen como ejes del cubo y las métricas se obten-

drían a partir de la combinación de dos o más dimensiones, apareciendo el resultado sobre la cara del cubo que corresponda a su intersección.

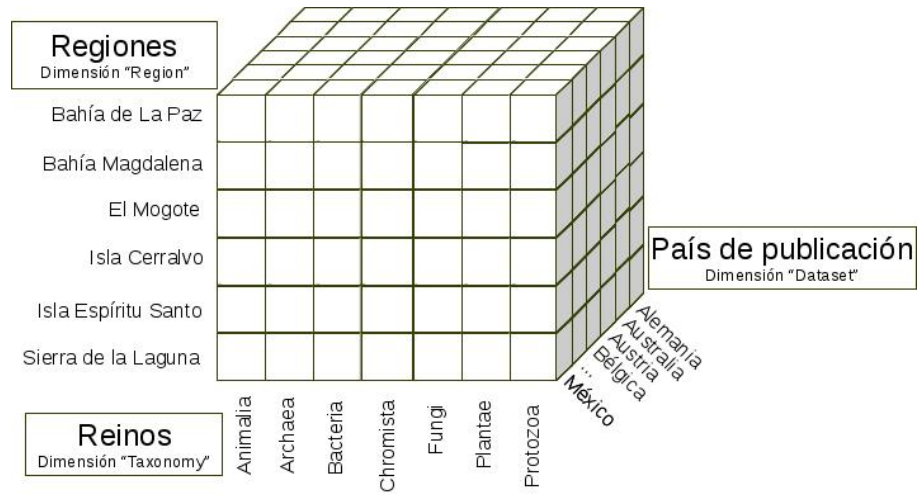


Figura 3.22: Representación gráfica del cubo que muestra las diferentes dimensiones del modelo. En este aparecen señaladas las dimensiones “Region”, “Taxonomy” (en su nivel reino) y “Dataset” (en su nivel país de publicación).

### 3.3.1. Una consulta en dos dimensiones

Para la primer consulta elegimos dos dimensiones: la región y el reino, buscando obtener el número de ocurrencias para las combinaciones de estos atributos. La consulta MDX para obtener este cálculo sería:

```
SELECT
  {[Taxonomy].[Kingdom].[Kingdom]} ON COLUMNS,
  {[Region].[Region Name].[Region Name]} ON ROWS
FROM [biodiversity]
```

El resultado de este cálculo podemos verlo en la fig. 3.23, en forma de una matriz.

Con esta consulta podemos saber la cantidad de organismos que, para cada grupo a nivel reino, se encontraron en cada una de las regiones definidas.

	Animalia	Archaea	Bacteria	Chromista	Fungi	Plantae	Protozoa
Bahía de La Paz	8337	(null)	(null)	640	2	2008	176
Bahía Magdalena	5086	(null)	(null)	681	151	455	323
El Mogote	370	(null)	(null)	40	(null)	479	3
Isla Cerralvo	2421	(null)	(null)	(null)	(null)	1020	(null)
Isla Espíritu Santo	4969	(null)	(null)	(null)	(null)	1293	(null)
Sierra de la Laguna	12303	(null)	(null)	(null)	812	6437	(null)

Figura 3.23: Resultado obtenido de la consulta MDX para reinos y regiones, en donde las etiquetas verticales a la izquierda representan los valores del atributo [Region Name] de la dimensión “Region” y las etiquetas horizontales sobre la matriz representan los valores del atributo [Kingdom] de la dimensión “Taxonomy”.

Este resultado en forma matricial sería equivalente a explorar gráficamente una de las caras del cubo multidimensional. Así, la intersección de los valores de las dimensiones “Region” y “Taxonomy” da como resultado una métrica o indicador, quedando estas métricas sobre la cara del cubo en la que convergen, tal y como se muestra en la fig. 3.24.

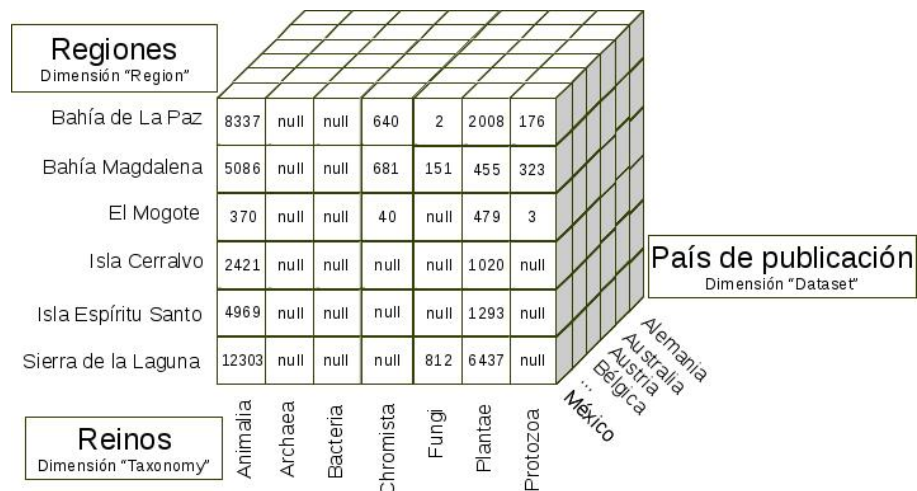


Figura 3.24: Representación gráfica del cubo que muestra las métricas obtenidas al combinar las dimensiones “Region” y “Taxonomy”.

### 3.3.2. Rotando el cubo

Si deseamos obtener las métricas a partir de otras dimensiones, basta con cambiarlas en la selección de las columnas y renglones dentro de la consulta MDX. Por ejemplo, para obtener la



métricas de los países de publicación de los registros por reino, utilizaríamos la consulta:

```
SELECT
{[Taxonomy].[Kingdom].[Kingdom]} ON COLUMNS,
{FILTER([Dataset].[Country Name].[Country Name],
[Measures].[Occurrence Count] <> NULL)} ON ROWS
FROM [biodiversity]
```

En donde, además, se añadió un filtro para descartar métricas vacías y así reducir el número de renglones visualizados. Esta consulta da como resultado la matriz mostrada en la fig. 3.25.

	Animalia	Archaea	Bacteria	Chromista	Fungi	Plantae	Protozoa
Alemania	(null)	(null)	(null)	93	41	3	(null)
Australia	6	(null)	(null)	(null)	22	4	(null)
Austria	2	(null)	(null)	(null)	20	15	(null)
Bélgica	33	(null)	(null)	(null)	(null)	(null)	(null)
Brasil	3	(null)	(null)	(null)	(null)	5	(null)
Canadá	4416	(null)	(null)	(null)	10	111	(null)
Colombia	(null)	(null)	(null)	(null)	(null)	1	(null)
Costa Rica	(null)	(null)	(null)	(null)	(null)	1	(null)
España	(null)	(null)	(null)	(null)	(null)	3	(null)
Estados Unidos	407874	(null)	(null)	19	11349	10134	10
Francia	186	(null)	(null)	(null)	(null)	(null)	(null)
Japón	7	(null)	(null)	(null)	(null)	(null)	(null)
México	68265	(null)	776	4703	15	90462	2552

Figura 3.25: Resultado obtenido de la consulta MDX para reinos y países de publicación.

Esta operación de consulta sería equivalente a rotar el cubo multidimensional, dejando de frente la cara con la combinación de las dimensiones “Taxonomy” y “Dataset”, en sus niveles reino y país de publicación, respectivamente. Esta representación gráfica aparece en la fig. 3.26.

### 3.3.3. Descendiendo en la jerarquía

Podemos, además, consultar a mayor detalle la información de un atributo particular en una dimensión, cuando esta posee una jerarquía definida. Por ejemplo, para obtener las métricas de los filios del reino Animalia en las diferentes regiones, utilizaríamos la consulta:

```
SELECT
```

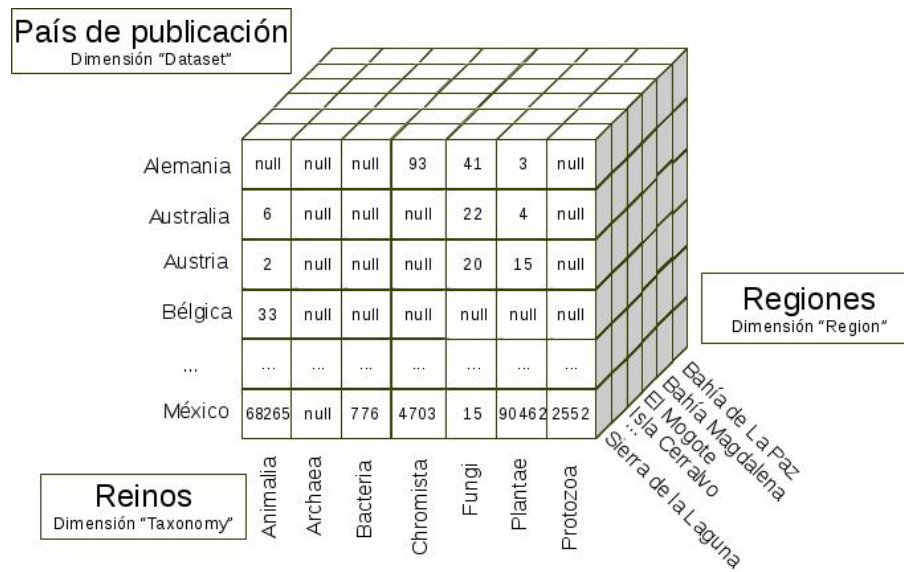


Figura 3.26: Representación gráfica del cubo que muestra las métricas obtenidas al combinar las dimensiones “Dataset” y “Taxonomy”.

```
[ Taxonomy ] . [ Phylum ] . [ Phylum ] ON COLUMNS,
[ Region ] . [ Region Name ] . [ Region Name ] ON ROWS
FROM [ biodiversity ]
WHERE [ Taxonomy ] . [ Kingdom ] . & [ Animalia ]
```

El resultado de esta consulta se muestra en la fig. 3.27.

	Acanthocephala	Annelida	Arthropoda	Brachiopoda	Bryozoa	Chordata	Cnidaria	Echinodermata	Mollusca
Bahía de La Paz	12	18	101	(null)	(null)	7799	87	99	144
Bahía Magdalena	(null)	24	70	(null)	(null)	4891	3	39	56
El Mogote	(null)	(null)	49	(null)	(null)	289	16	14	2
Isla Cerralvo	1	1	71	(null)	(null)	2194	117	5	16
Isla Espíritu Santo	(null)	(null)	52	(null)	(null)	4365	119	261	119
Sierra de la Laguna	(null)	(null)	428	(null)	(null)	11868	(null)	(null)	(null)

Figura 3.27: Resultado obtenido de la consulta MDX para los filos del reino Animalia por regiones geográficas, en donde se añadió una condición para acceder a los filos de este reino en particular.

De esta forma, basta con seleccionar una atributo particular en una dimensión para explorarlo a mayor detalle, descendiendo dentro de su jerarquía. Esta operación de selección sería equivalente a extraer una columna particular del cubo mutidimensional para explorar su jerarquía, como se ilustra en la fig. 3.28.

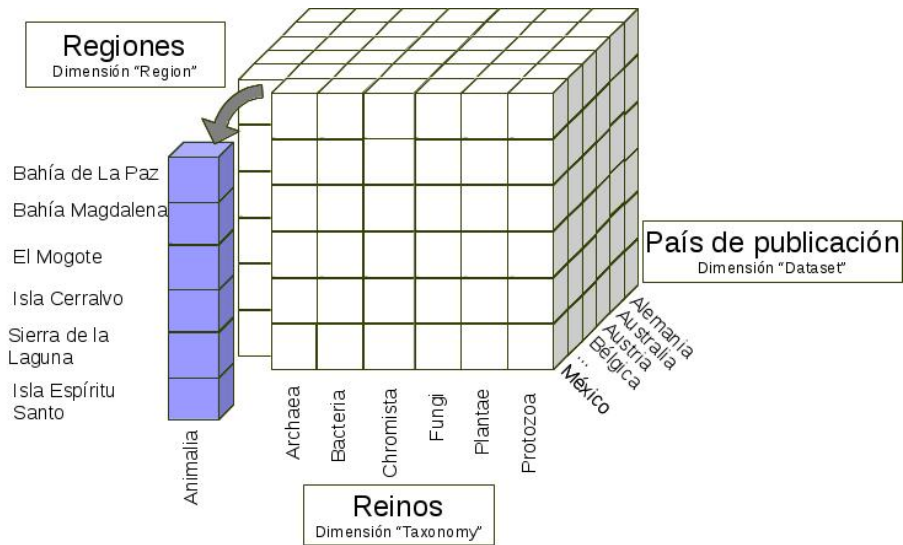


Figura 3.28: Representación gráfica de la selección del reino Animalia dentro del cubo multidimensional para explorar la jerarquía de la dimensión "Taxonomy".

### 3.3.4. Una consulta en tres dimensiones

Las consultas, además, puedan involucrar más de dos dimensiones, de forma que se obtenga un conjunto de métricas que responda a preguntas más específicas. Por ejemplo, si deseamos conocer los reinos y el país de publicación de los especímenes colectados en la Bahía de La Paz, tendríamos que utilizar tres dimensiones distintas: "Taxonomy", "Dataset" y "Region". La consulta que arrojaría esta información sería:

```
SELECT
  {[Taxonomy].[Kingdom].[Kingdom]} ON COLUMNS,
  {FILTER([Dataset].[Country Name].[Country Name],
  [Measures].[Occurrence Count] <> NULL)} ON ROWS
FROM [biodiversity]
WHERE [Region].[Region Name].&[Bahia de La Paz]
```

Arrojando como resultado la matriz mostrada en la fig. 3.29.

Esta consulta se representaría gráficamente como se visualiza en la fig. 3.30.

	Animalia	Archaea	Bacteria	Chromista	Fungi	Plantae	Protozoa
Canadá	4	(null)	(null)	(null)	(null)	(null)	(null)
Estados Unidos	6946	(null)	(null)	6	(null)	26	(null)
Francia	5	(null)	(null)	(null)	(null)	(null)	(null)
México	525	(null)	(null)	634	2	1982	176
ND	855	(null)	(null)	(null)	(null)	(null)	(null)
Países Bajos, Holanda	2	(null)	(null)	(null)	(null)	(null)	(null)

Figura 3.29: Resultado obtenido de la consulta MDX para obtener el reino y país de publicación, añadiendo en la condición que se obtengan únicamente los registros localizados en la Bahía de La Paz.

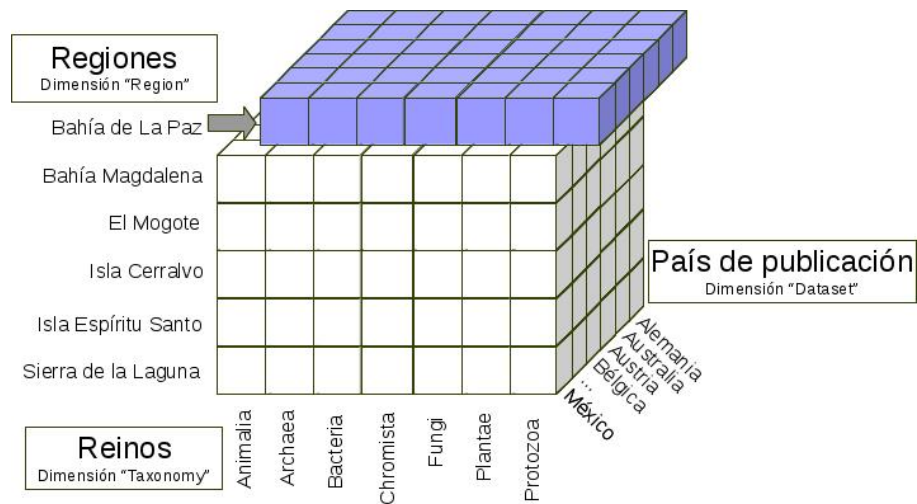


Figura 3.30: Representación gráfica de la consulta de las dimensiones "Taxonomy", "Dataset" y "Region", seleccionando únicamente las ocurrencias de la Bahía de La Paz.

### 3.3.5. Añadiendo la dimensión del tiempo

Finalmente, para complementar las consultas requeridas para el análisis de la información, se necesitó añadir en éstas la dimensión “Time”, para así obtener una perspectiva mucho más completa de la información, al revisar como han variado las métricas con el paso del tiempo.

Para lograrlo bastó tomar cualquiera de las consultas antes vista y añadir un atributo de la dimensión “Time” en alguno de los tres elementos clave de la consulta: columnas, renglones o condiciones.

Por ejemplo, para obtener el reino de los organismos registrados en el periodo comprendido entre 2003 y 2013, se ejecutaría la consulta:

```
SELECT
{[Taxonomy].[Kingdom].[Kingdom]} ON COLUMNS,
{FILTER([Time].[Year].[Year].[Calendar 2003]:
[Time].[Year].[Year].[Calendar 2013],
[Measures].[Occurrence Count] <> NULL)} ON ROWS
FROM [biodiversity]
```

Dando como resultado la matriz de métricas que se muestra en la fig. 3.31.

	Animalia	Archaea	Bacteria	Chromista	Fungi	Plantae	Protozoa
Calendar 2003	16271	(null)	155	17	12	339	20
Calendar 2004	13412	(null)	438	(null)	(null)	110	(null)
Calendar 2005	18967	(null)	30	(null)	(null)	36	(null)
Calendar 2006	13398	(null)	(null)	(null)	(null)	79	(null)
Calendar 2007	13042	(null)	(null)	3	(null)	64	26
Calendar 2008	13987	42	370	(null)	(null)	96	(null)
Calendar 2009	25016	(null)	(null)	(null)	(null)	8	(null)
Calendar 2010	37322	(null)	(null)	(null)	(null)	14	(null)
Calendar 2011	25025	(null)	(null)	(null)	(null)	1	(null)
Calendar 2012	20387	(null)	(null)	(null)	(null)	19	(null)
Calendar 2013	34960	(null)	(null)	(null)	(null)	12	(null)

Figura 3.31: Resultado obtenido de la consulta MDX para obtener el reino de los organismos registrados en el periodo entre 2003 y 2013. En esta consulta se añade un rango de fechas en la selección de los renglones, para así seleccionar solo los años que nos interesa analizar.

Esta consulta se representaría gráficamente como se visualiza en la fig. 3.32, donde se aprecia que la dimensión temporal es transversal al resto de las dimensiones, permitiéndonos obtener información sobre el cambio en las métricas desde distintas perspectivas a través del tiempo.

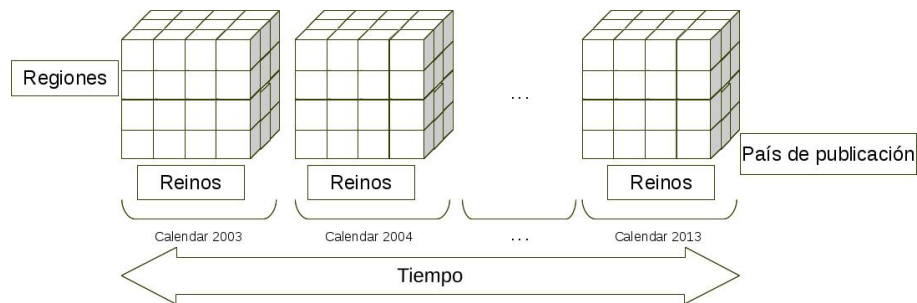


Figura 3.32: Representación gráfica de la consulta de las dimensiones “Taxonomy” y “Time”, en un rango que va desde el año 2003 hasta 2013. En esta figura se representa a la dimensión temporal de forma transversal al resto de las dimensiones antes vista, ya que es posible combinar cualquiera de las otras dimensiones del modelo con la dimensión “Time” y analizar todas sus métricas desde una perspectiva temporal.

## Capítulo 4

# Construcción del cliente OLAP móvil

El modelo multidimensional implementado permite analizar la información de biodiversidad desde distintas perspectivas y responder a preguntas relevantes para los investigadores y estudiantes de las áreas de biología y ecología. Sin embargo, esta implementación del modelo estaba limitada a la interfaz de consulta que brinda el software utilizado en el equipo donde se desarrolló el hipercubo, limitando así las posibilidades de utilizarlo entre los usuarios que llevarían a cabo el análisis. Por este motivo, se construyó una aplicación especializada que facilitara el análisis multidimensional de los datos en una arquitectura cliente - servidor, de modo que fuera posible a varios usuarios conectarse simultáneamente al sistema y analizar las métricas del hipercubo. Además, en esta aplicación se aprovecharían también la información primaria sobre biodiversidad obtenida de los servicios web de acceso abierto ya disponibles, de modo que se tuviera una visión más integral de la información.

Se eligió desarrollar esta aplicación en un entorno de dispositivos móviles, debido a las necesidades particulares de movilidad que encontramos entre los usuarios a los que está destinado este sistema. En la fig. 4.1 se especifican los servicios web involucrados y el papel que jugará la aplicación móvil.

Este desarrollo se enfocó en satisfacer los tres requerimientos principales, detectados durante la investigación documental del problema de estudio, los cuales son:

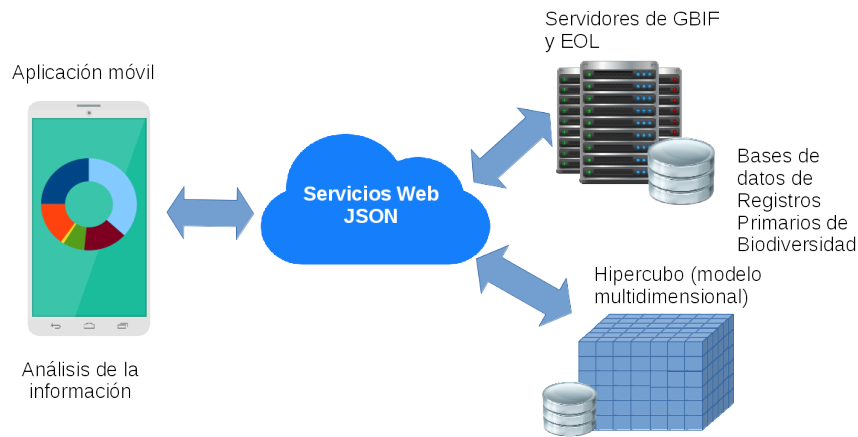


Figura 4.1: Diagrama de la comunicación entre la aplicación móvil y los servicios web explotados. En este diseño, el cliente móvil se utiliza como la puerta de acceso a la información obtenida en las bases de datos de acceso abierto y las consultas multidimensionales del hipercono.

1. Construir un mapa unificado global de la biodiversidad fácilmente accesible.
2. Compartir y sintetizar rápidamente estos datos y los conocimientos que nos proporcionan.
3. Visualizar y analizar los datos en múltiples dimensiones.

El primer paso consistió en lograr la visualización de los datos sobre biodiversidad en un mapa global unificado, explotando los servicios web de información georreferenciada de acceso abierto de la GBIF y la EOL, para satisfacer el primer punto. Para atender los requerimientos de los puntos dos y tres, el siguiente paso fue explotar el servicios web de información estadística global de la GBIF, junto con la información multidimensional proporcionada por el hipercono diseñado.

## 4.1. Consulta y visualización de un mapa sobre biodiversidad

La aplicación desarrollada se programó en el lenguaje Java, empleando para su prueba un emulador de dispositivos móviles Android. Dentro de su desarrollo se incluyó un mapa interactivo para consultar la información de las ocurrencias de organismos y las especies en una región



geográfica. Los mapas son proporcionados por la API del proyecto OpenStreetMap, gracias a la cual el usuario tiene la posibilidad de interactuar con el mapa, seleccionando el punto geográfico que se desea analizar, y cargando las ocurrencias de ejemplares en forma de etiquetas sensibles a su selección por medio de una interfaz táctil.

La información sobre los PBR y del catálogo de especies se obtuvieron explotando los servicios web en formato JSON de la GBIF y la EOL, respectivamente. Se utilizó el servicio web de consulta de ocurrencias de especímenes de la GBIF, para obtener los PBR georreferenciados de esta base de datos, que se encontraran dentro de el área seleccionada por el usuario. Mientras que el servicio de la EOL, nos proporcionó la información complementaria sobre las especies, que incluyen fotografías, listado de nombres comunes y descripciones.

Debido a la gran cantidad de información que se puede recuperar de una área geográfica, las consultas se limitaron a no poder traer más de 300 registros en cada llamada al servicio, dejando en manos del usuario especificar este límite en un rango de 30 a 300. Además, dado que el servicio web no puede validar si los registros obtenidos en cada llamada fueron devueltos previamente al cliente, se aplicó contenido algorítmico para evitar la duplicidad de los datos y mejorar la velocidad de respuesta, utilizando estructuras de listas ligadas para almacenar la información de las ocurrencias en memoria, y una implementación del algoritmo de búsqueda binaria. De esta forma, aunque la lista de ocurrencias crecía rápidamente formando un conjuntos de miles de registros, el costo del proceso de verificación de no duplicidades se mantenía en un tiempo  $O(\lg n)$  Cormen et al. (2009). Junto a estos algoritmos, se empleó cómputo paralelo a través de múltiples hilos de ejecución, con el objetivo de mejorar el rendimiento de la aplicación, al ejecutar varias llamadas simultaneas a diferentes servicios web, durante la consulta de la información detallada de las especies.

La apariencia de este mapa interactivo y la forma de consultar la información, se muestra en la fig. 4.2.

Además de las consultas de las ocurrencias georreferenciadas, se añadieron diferentes funciones para consultar información estadística sobre las ocurrencias de especímenes.

La primera función estadística consistió en una visualización del número de especies y es-

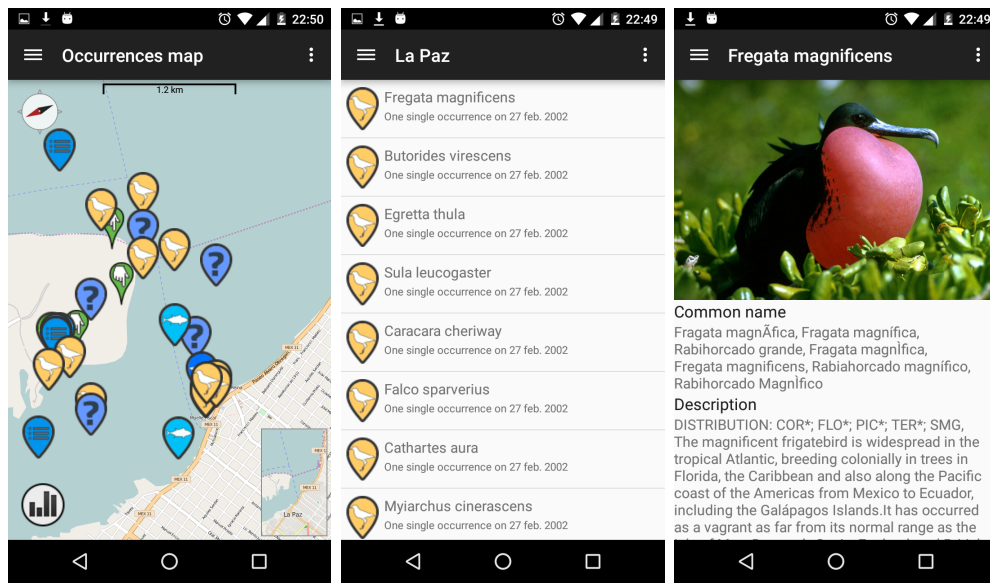


Figura 4.2: Captura de la interfaz de la aplicación móvil, que muestra el mapa interactivo para consultar las ocurrencias de ejemplares en una región. En éste es posible seleccionar un punto geográfico para buscar a su alrededor ocurrencias de especímenes registrados en la base de datos de la GBIF.

pecímenes encontrados en el mapa, desde las dimensiones de su clasificación taxonómica y su año de captura. Se obtuvo una vista como la que se muestra en la fig. 4.3.

Esta restricción en cuanto a la cantidad de dimensiones se debió a las limitaciones de almacenamiento en los dispositivos móviles, ya que la información estadística es calculada al momento de realizar las consultas sobre el mapa interactivo, y se almacena en la memoria interna del dispositivo dentro de archivos de texto, por lo que fue necesario restringir el número de dimensiones para limitar el tamaño final de los archivos.

La segunda función estadística consistió en un módulo para acceder a las sumatorias de ocurrencias a nivel mundial de la GBIF. En este módulo, solo se pueden seleccionar las dimensiones de base del registro, país de ocurrencia, país de publicación y los años del registro, teniendo únicamente la opción de combinar las dimensiones de país de ocurrencia y país de publicación para obtener una vista filtrada de la información, tal como se muestra en la fig. 4.4.

Estas limitaciones se deben a las características de los servicios web proporcionados por la GBIF, que no permiten obtener una visualización de la información en la que se combinen dos

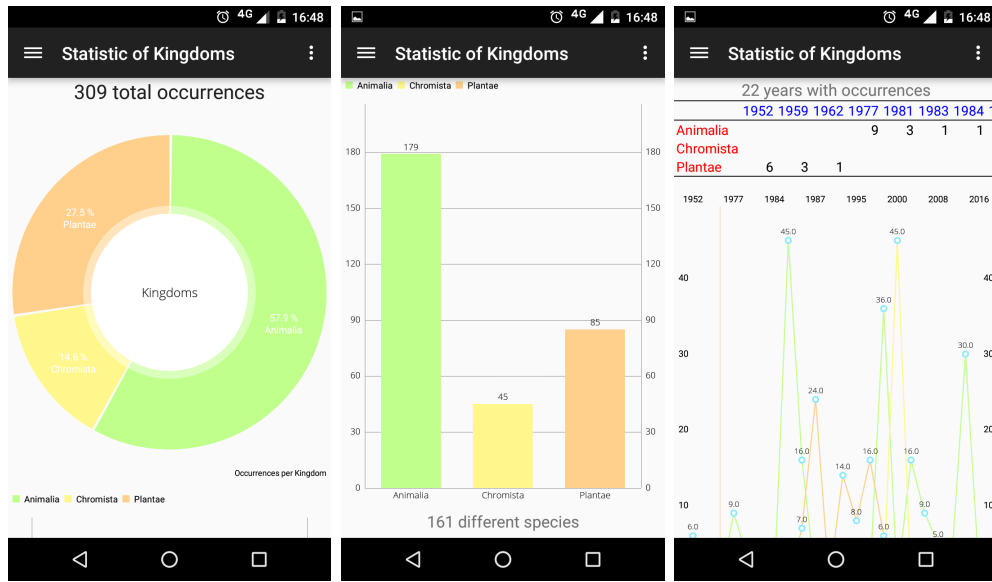


Figura 4.3: Captura de la interfaz de la aplicación móvil, que muestra la información estadística sobre ocurrencias consultadas a través del mapa interactivo.

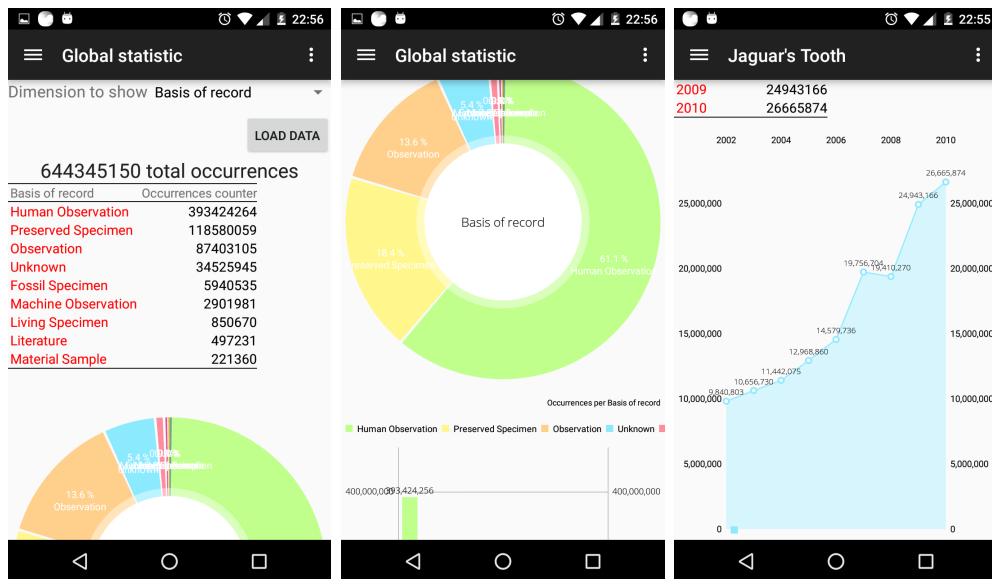


Figura 4.4: Captura de la interfaz de la aplicación móvil, que muestra la información estadística global sobre ocurrencias extraída de la base de datos centralizada de la GBIF.

o mas dimensiones, razón por la cual, sumado a las limitaciones de las estadísticas del mapa mencionadas antes, fue necesario desarrollar una función adicional en la aplicación móvil, para explotar la información obtenida del modelo multidimensional implementado.

## 4.2. Servicio web para la consulta del hipercubo

Para que el cliente móvil pudiera acceder a la información del hipercubo, se creó un servicio web basado en la tecnología ASP.Net MVC, utilizando el formato JSON para el intercambio de los datos. Este servicio se ejecutó sobre un servidor web IIS, junto con el servidor de consultas de SQL Server Analysis Services, de modo que el servicio funciona como un intermediario entre el cliente y el hipercubo (ver fig. 4.5), haciendo el trabajo de ejecutar las consultas multidimensionales y devolver la información obtenida.

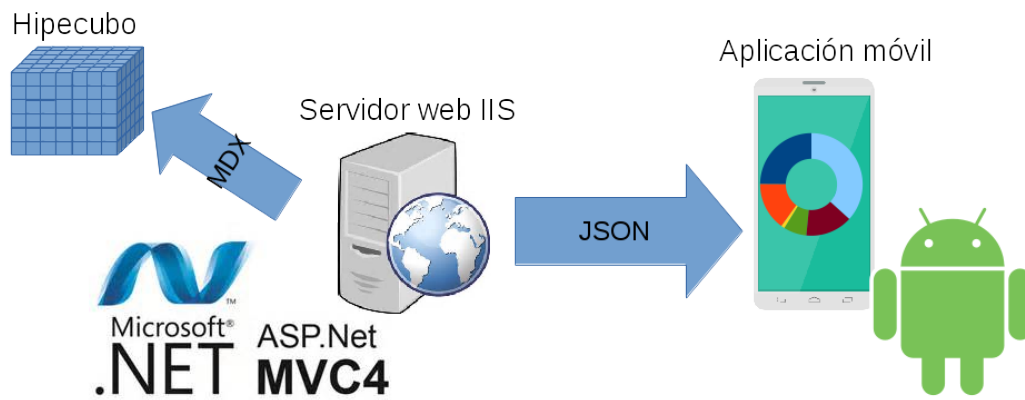


Figura 4.5: Diagrama de la comunicación entre el cliente móvil y el hipercubo. En este diseño, el servicio web funciona como un intermediario entre el cliente y el hipercubo, encargándose de construir las consultas multidimensionales a partir de los parámetros recibidos, liberando al cliente de todos los detalles sobre la elaboración de las consultas MDX

Para la comunicación entre el cliente y el servicio se utilizan los parámetros contenidos en la URL de la petición HTTP, lo cuales determinan la función que ejecutará el servicio web, para finalmente traducir los parámetros en una instrucción MDX que es ejecutada por el hipercubo. De esta forma, el servicio brinda cuatro funciones básicas:

1. **Name.** Función que devuelve el nombre del hipercubo consultado, utilizada únicamente

para identificar al servicio web del lado del cliente y probar que la comunicación funciona adecuadamente.

2. **Dimensions.** Devuelve una lista con los nombres de cada una de las dimensiones del hipercubo, junto con la lista de los valores de los atributos del primer nivel de cada una de sus jerarquías.
3. **Dimension.** Devuelve una lista de atributos en el nivel de la jerarquía especificado por los parámetros contenidos en la URL, lo cual es útil cuando deseamos descender dentro de la jerarquía de una dimensión, y saber cuales atributos se encuentran dentro de ese nuevo nivel.
4. **Occurrences.** Esta es la función más importante del servicio web, ya que al utilizarla podemos ejecutar las consultas multidimensionales diseñadas. Recibe como parámetros las dimensiones que se desean visualizar y el filtro que se utilizará para hacer una consulta con mayor detalle, devolviendo el resultado dentro de las etiquetas JSON de la respuesta, en un formato de texto plano que simula una matriz, similar a la obtenida por la herramienta de consulta de SQL Server Analysis Services.

Una vez programado el servicio web, se probó a través del navegador web, enviando diferentes parámetros que correspondieran a las consultas multidimensionales ya implementadas con anterioridad, tal y como se muestra en la fig. 4.6.

De esta forma, cualquier aplicación tipo cliente adecuadamente programada, puede conectarse con el hipercubo y ejecutar las consultas multidimensionales diseñadas.

### **4.3. Interfaz de consulta multidimensional**

Una vez construido el servicio web, se programó una función de consulta en la aplicación móvil, de tal manera que la información proporcionada por el hipercubo pudiera ser explotada en este medio, obteniendo de esta forma un cliente OLAP móvil.

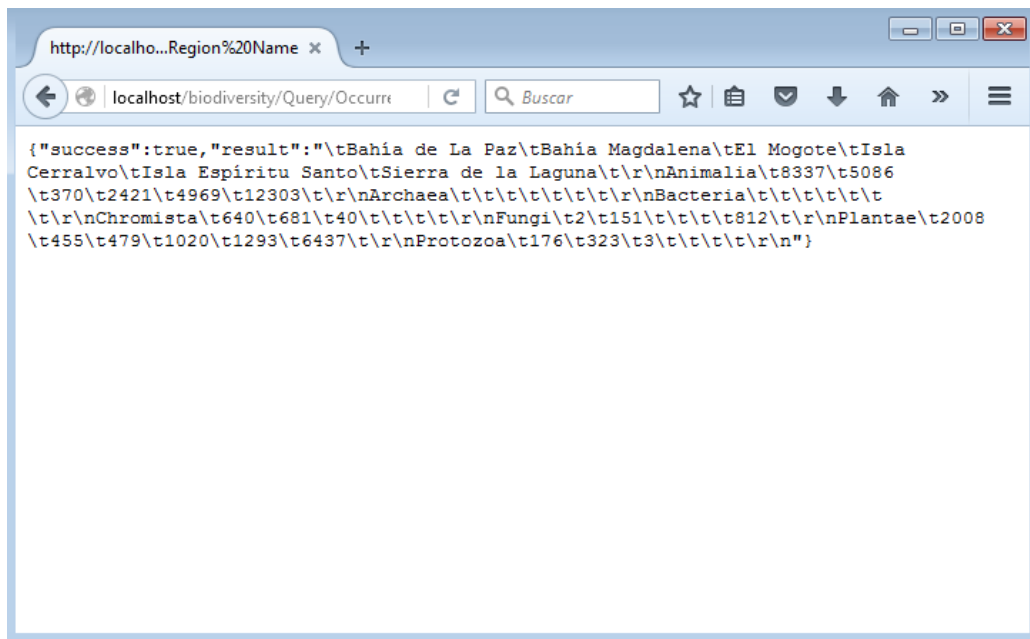


Figura 4.6: Captura de la prueba del servicio web a través de un navegador de Internet. En esta llamada al servicio se ejecuta la función Occurrences, solicitando una consulta de las dimensiones “Taxonomy” y “Region”, en sus niveles “Kingdom” y “Region Name”, respectivamente.

La ejecución de las consultas multidimensionales se logran por medio de un formulario diseñado para definir los parámetros para el servicio web.

En este formulario, el usuario selecciona las dimensiones que desea analizar para, finalmente, visualizar las métricas obtenidas por medio de tablas y distintos tipos de gráficos. Es posible utilizar hasta tres dimensiones en cada consulta; las primeras dos, especificadas en el formulario por los campos `Dimension for rows` y `Dimension for columns` (ver fig. 4.7), determinan las dimensiones que se incluirán dentro de la información visualizada en la tabla de resultados, seleccionando primero las dimensiones a analizar. Posteriormente se eligen los atributos que se incluirán en la consulta y que serán visualizados en los renglones y columnas del resultado. Además, es posible descender dentro de la jerarquía de estas dos dimensiones, al seleccionar los atributos a visualizar, ya que estos se muestran como una lista ordenada del nivel más alto hasta el más bajo. Finalmente, se puede incluir una tercera dimensión en la consulta, que funciona como un filtro, especificada por el campo `Filter`, la cual permite condicionar la consulta a un valor específico de los atributos de la dimensión seleccionada. De este modo podemos obtener respuestas para preguntas mucho más específicas, tal y como se verá en el capítulo Resultados.

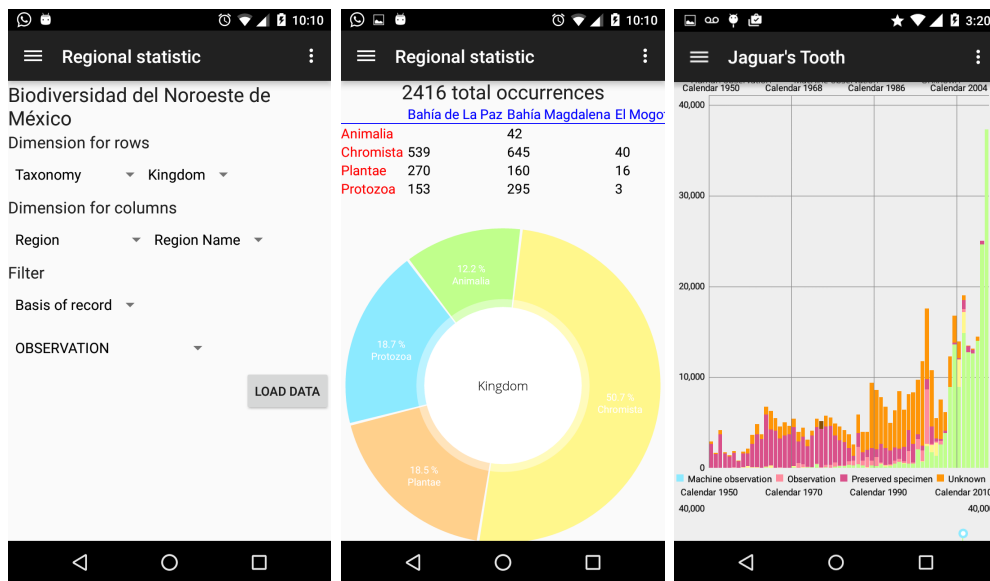


Figura 4.7: Captura de la interfaz de la aplicación móvil, que muestra el formulario para las consultas multidimensionales. En este formulario, es posible combinar hasta tres dos dimensiones para las consultas (dos para los renglones y columnas de la matriz resultante y una tercera que funciona como filtro).

# Capítulo 5

## Resultados

Como producto final de este proyecto podemos listar como resultados más importantes:

- Se brinda al usuario un sistema de consulta de la información primaria sobre biodiversidad, basado en un mapa interactivo de ocurrencias de especímenes, con el cual el usuario puede visualizar la variedad de organismos encontrados en una región geográfica específica y acceder a la información biológica complementaria, de las distintas especies que encuentre durante su búsqueda. De este modo, el usuario puede acceder directamente a la información georreferenciada de la base de datos de la GBIF, basándose en una posición geográfica, sin la necesidad de formularios de captura. Los datos consultados se almacenan en la memoria interna del dispositivo móvil, para consultarlos posteriormente fuera de línea. Así, los investigadores pueden tener disponible en todo momento la información consultada al realizar su trabajo de campo, sabiendo qué especies podrán encontrar en la región de estudio y la información complementaria de estas especies, que incluyen fotografías, descripción y listado de nombres comunes.
- Se obtuvo un modelo de datos multidimensional que permite analizar la información sobre biodiversidad, a través de métricas vistas desde varias perspectivas y con diferentes niveles de detalle. Este modelo fue probado con un fragmento obtenido de la base de datos de PBR de la GBIF, de forma que puede ser extrapolado a conjuntos de datos mucho más grandes que abarquen áreas geográficas mayores, permitiéndonos realizar un análisis mucho



más amplio sobre la biodiversidad de las regiones que nos interesen. Gracias a este modelo, los usuarios de la información sobre biodiversidad pueden contestar preguntas complejas y relevantes para su trabajo y la toma de decisiones, como podrían ser:

- ¿Qué país posee el mayor inventario de capturas de especímenes realizadas en territorio mexicano?
  - ¿A qué bases de datos podemos acudir para obtener la información sobre las familias de plantas con flores de la región?
  - ¿En qué años hubo mayor número de registros de peces en Bahía Magdalena?
  - ¿Qué región del Noroeste de México cuenta con la mayor presencia de lobos marinos?
  - ¿Cómo ha variado la forma de registrar la biodiversidad a los largo del tiempo?
- El servicio web desarrollado y a la interfaz gráfica incorporada en la aplicación móvil, permiten al usuario acceder rápidamente a la información resumida sobre la biodiversidad contenida en el hipercubo, debido a que este utiliza un modelo multidimensional en el que se almacenan únicamente las dimensiones diseñadas y sus combinaciones para obtener distintas métricas, en lugar de recurrir a consultas SQL estándar, que están forzadas a leer cada uno de los cientos de miles de registros de las tablas de la base de datos, para cada cálculo requerido. Por ejemplo, una consulta SQL para obtener la cantidad de ocurrencias de cada especie tarda hasta 24 segundos, mientras que este mismo cálculo utilizando una consulta MDX en el hipercubo consume sólo 2 segundos. Además, como la información ya fue sintetizada por el modelo implementado, pudo ser adaptada para visualizarla en la pantalla de los dispositivos móviles, de modo que resulta fácil para el usuario comprender los resultados entregados en la palma de su mano. Esta información, al igual que la que es extraída de las bases de datos de la GBIF y la EOL, es almacenada en la memoria interna del dispositivo, de modo que también puede ser consultada posteriormente sin la necesidad de un nuevo acceso a Internet.

## 5.1. Funcionalidades del cliente OLAP móvil

Gracias a su interacción con los servicios web de las bases de datos globales de biodiversidad, con el servicio web de consultas multidimensionales, y a sus capacidades de movilidad, el cliente OLAP móvil posee las siguientes funcionalidades:

1. El mapa interactivo para consulta de ocurrencias, donde a través de una interfaz táctil, el usuario puede seleccionar un punto geográfico en el mapa y consultar las ocurrencias localizadas en un radio aproximado de 2 km. Dentro de este mapa interactivo, los organismos localizados en un mismo punto geográfico (misma latitud y longitud), se agrupan bajo una misma etiqueta que también puede ser seleccionada por el usuario, para desplegar la lista de especies localizadas en ese punto, distinguiendo los elementos de la lista por sus nombres científicos y un icono representativo de la clase a la que pertenecen (mamíferos, aves, reptiles, insectos, etc.), junto con la fecha de las ocurrencias y el número de organismos registrados. Para llegar a un mayor nivel de detalle, se puede seleccionar un elemento de la lista y acceder al resto de los datos de la ocurrencia, incluyendo la información complementaria de la especie (imágenes, nombres comunes y descripciones), el tipo de registro realizado (observación, espécimen preservado, etc.) y el nombre del proyecto o colección biológica que posee el registro. De esta forma, el usuario obtiene acceso a la información primaria sobre biodiversidad, de las ocurrencias de organismos registradas en la GBIF, así como a la información complementaria de las especies obtenida de la EOL.
2. La consulta de estadísticas del mapa, que muestra información estadística sobre el número de ocurrencias obtenidas en el mapa interactivo, permitiendo analizar la información de acuerdo a la clasificación taxonómica y al año de cada registro.
3. La función de consulta estadística global, con la cual se obtiene acceso a la información estadística sobre ocurrencias de la base de datos centralizada de la GBIF. En esta función, por medio de un formulario se obtienen estadísticas para ser visualizadas en tablas y gráficas, que incluyen la cantidades globales de ocurrencias diferenciadas por las dimensiones soportadas por el servicio web de la GBIF: base del registro, años, país de la ocurrencia y país de publicación.

4. La función de consulta multidimensional, que brinda acceso a las métricas sobre ocurrencias del hipercubo regional desde múltiples dimensiones. Para esta función, también se emplea un formulario para ejecutar las consultas, pero con la particularidad de que se pueden combinar hasta tres dimensiones en una misma consulta, dando al usuario la posibilidad de elegir entre seis diferentes perspectivas para hacer estas combinaciones: grupo de organismos, región, lugar, base del registro, colección y tiempo. Para la selección, se eligen los atributos de las dimensiones que se desean analizar, pudiendo desplazarse entre las jerarquías de todas las dimensiones involucradas; las dos primeras serán las que se visualizarán en el resultado, mientras que la tercera dimensión funciona como un filtro, en el que podemos elegir un valor específico para el atributo seleccionado, y obtener un mayor nivel de detalle. La información obtenida es visualizada por medio de tablas y distintos tipos de gráficas, que se adaptan a la resolución y rotación de la pantalla del dispositivo, para facilitar así su comprensión. Gracias a la flexibilidad de esta interfaz gráfica y a la capacidad de las consultas MDX, es posible combinar las dimensiones de modo que se logre obtener conocimiento que de respuesta a preguntas complejas.
5. Acceso a la información fuera de línea, al utilizar el almacenamiento interno del dispositivo móvil. De esta forma, después de ejecutar cualquiera de las consultas antes descritas, la información obtenida es guardada en la memoria interna del dispositivo por medio de archivos de texto en formato JSON, de modo que después de suspender el dispositivo y volver a encenderlo, o al desconectarse de Internet, es posible seguir visualizando las ocurrencias sobre el mapa, así como volver a obtener la información de las especies y los datos estadísticos que ya hayan sido consultados.

## 5.2. Preguntas contestadas

Uno de los objetivos principales consistió en dar respuesta a preguntas complejas sobre biodiversidad, utilizando el modelo de datos diseñado. Este objetivo se logra al utilizar la función de consulta multidimensional del cliente móvil, e interpretar los resultados que nos arroja. Para probar esto, a continuación aparecen una serie de preguntas sobre la biodiversidad de la región analizada, y como estas son contestadas con la información obtenida del hipercubo.

### 5.2.1. ¿Qué país posee el mayor inventario de capturas de especímenes realizadas en territorio mexicano?

Dentro de las referencias bibliográficas de la justificación de este proyecto, Edwards et al. (2000) menciona que resulta contrastante que los países desarrollados posean la mayor parte de los registros electrónicos y físicos sobre la biodiversidad de nuestro planeta, mientras que la mayor diversidad de especies se encuentra en los países en vías de desarrollo. Con el modelo fue posible probar esta afirmación, seleccionando en el cliente la dimensión “Dataset” en su nivel “Publishing country”, con lo que se obtuvieron los resultados que se observan en la fig. 5.1.

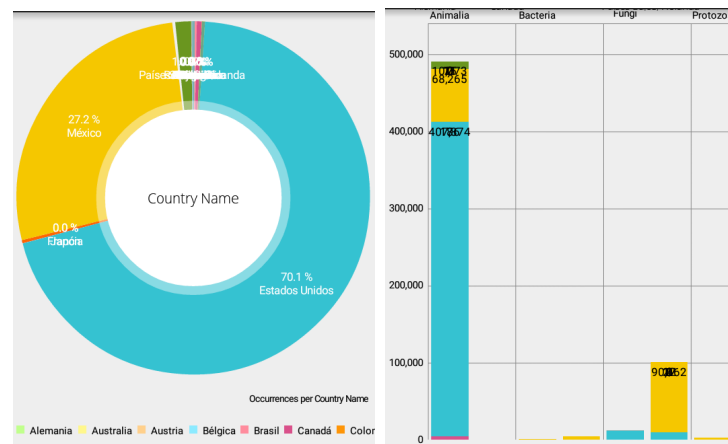


Figura 5.1: Captura de la interfaz de la aplicación móvil, que muestra la consulta de países de publicación de las ocurrencias dentro del territorio mexicano.

Como se puede ver, Estados Unidos posee más del 70% de los registros sobre biodiversidad obtenidos del territorio mexicano en su región noroeste, lo cual reforzaría la afirmación antes citada. Sin embargo, si añadimos una dimensión adicional a la consulta, pidiendo además distinguir la información por reinos con la dimensión “Taxonomy”, la perspectiva cambia, y como vemos en la segunda captura de la fig. 5.1, los registros de información sobre plantas se encuentran mayoritariamente en bases de datos nacionales.

Para poder dar respuesta plenamente a la pregunta original, solicitamos solo aquellas ocurrencias que cuenten con el ejemplar preservado por la organización que lo registró para, de esta forma, seleccionar sólo los registros que cuenten con un inventario físico de las capturas. Esto se logró agregando en el filtro la dimensión “Basis of record” con el atributo “Preserved specimen”,

obteniendo los resultados mostrados en la fig. 5.2.

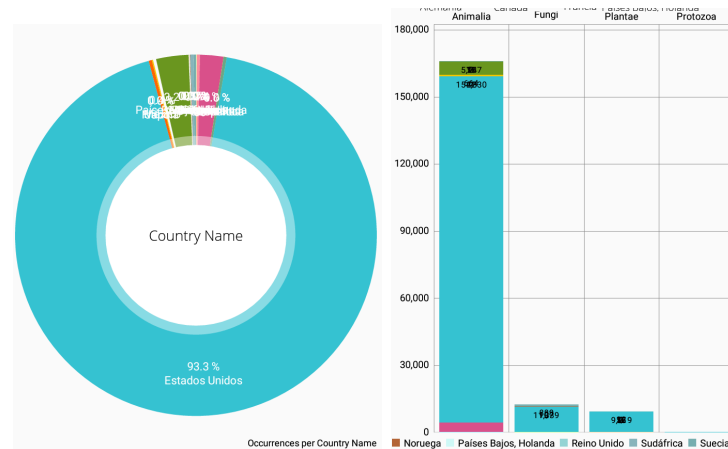


Figura 5.2: Captura de la interfaz de la aplicación móvil, que muestra la consulta de países de publicación de las ocurrencias dentro del territorio mexicano, que incluye sólo los registros de especímenes preservados.

Como podemos observar, Estados Unidos posee el 93.3 % de los registros de especímenes preservados sobre biodiversidad de la región, lo cual termina por dar respuesta a la pregunta planteada originalmente, además de confirmar lo señalado por los autores en el año 2000. Adicionalmente, este sería un conocimiento muy útil para un usuario que deseara encontrar datos sobre algún grupo de organismos en particular, ya que podría saber en que país sería más adecuado iniciar una búsqueda sobre las bases de datos que posean sus organizaciones, como se verá en la siguiente pregunta.

### 5.2.2. ¿A qué bases de datos podemos acudir para obtener la información sobre las familias de plantas con flores de la región?

Para contestar esta pregunta, utilizamos las dimensiones “Dataset” en su nivel “Dataset name” y “Taxonomy” en su nivel “Family”, en los renglones y columnas a mostrar, y aplicamos un filtro con la dimensión “Taxonomy”, para seccionar únicamente a las plantas con flores (que se encuentran dentro del orden Magnoliales). El resultado se puede visualizar en la fig. 5.3.

Llegando a la respuesta de que la base de datos del proyecto “Repatriación de datos del Herbario de Arizona (ARIZ)”, perteneciente a la CONABIO, posee la mayor cantidad de registros de

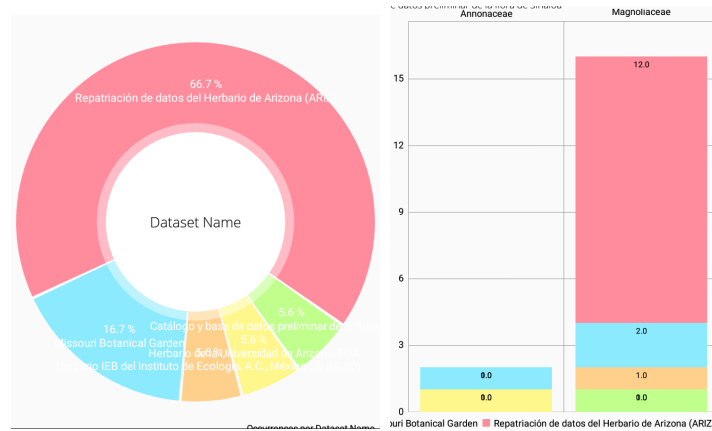


Figura 5.3: Captura de la interfaz de la aplicación móvil, que muestra la consulta de las bases de datos de las colecciones biológicas de plantas del orden Magnoliales, distinguiendo sus familias.

plantas con flores del orden Magnoliales, específicamente los pertenecientes a la familia de plantas Magnoliaceae, y que esta sería la base de datos a elegir para buscar información sobre este grupo de organismos.

### 5.2.3. ¿En qué años hubo mayor número de registros de peces en Bahía Magdalena?

Esta pregunta se responde combinando tres dimensiones: primero la dimensión “Region” en su único nivel “Region Name” para los renglones, seguida de la dimensión “Time” en su nivel “Year” para las columnas, especificando un rango desde 1950 hasta 2013, y finalizando con un filtro para incluir únicamente a los peces de la clase “Actinoptergii”. El resultado obtenido se muestra en la fig. 5.4.

Como se puede observar, los años 1971 y 1974 presentan las mayores ocurrencias de peces en Bahía Magdalena, lo cual es un indicador importante sobre la actividad de investigación en las pesquerías de esta región, de modo que un estudiante tendría mayores oportunidades de encontrar información sobre las especies de peces de su interés en esta región, localizando las publicaciones realizadas en estos años.

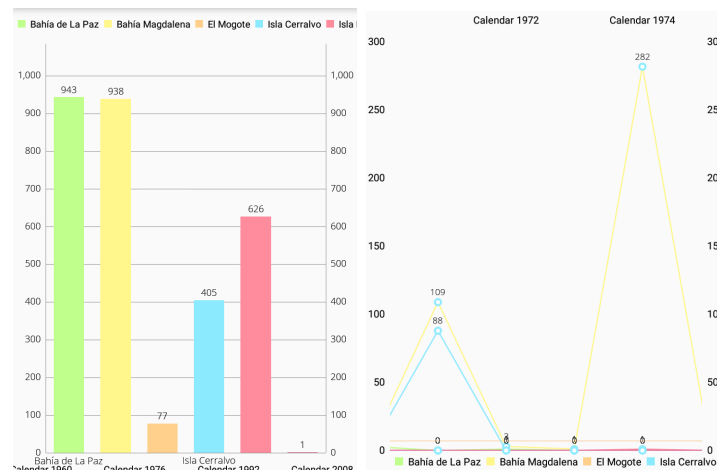


Figura 5.4: Captura de la interfaz de la aplicación móvil, que muestra las ocurrencias de peces de la clase “Actinoptergii” en las distintas regiones, dividiéndolas por años. En esta gráfica, Bahía Magdalena aparece señalada por el color amarillo.

#### 5.2.4. ¿Qué región del Noroeste de México cuenta con la mayor presencia de lobos marinos?

Las especies que llamamos comúnmente lobos marinos, son mamíferos que pertenecen a la familia Otariidae, por lo que la dimensión del filtro se referirá a esta clasificación taxonómica en particular, mientras que para los renglones, seleccionaremos la dimensión “Region” para mostrar los nombres de las regiones donde aparezcan ejemplares de esta familia. Adicionalmente, se añade en las columnas la dimensión “Taxonomy” en su nivel “Genus”, para distinguir los distintos géneros de lobos marinos que se encontraron. El resultado se aprecia en la fig. 5.5. en donde vemos que, dentro de las regiones del modelo, Bahía Magdalena es el lugar con mayor presencia de lobos marinos (familia Otariidae) con el 70 %, encontrando en esta a las especies pertenecientes al género *Zalophus*, mientras que Isla Espíritu Santo le seguiría con el 15 % de las ocurrencias, contando con especies de los géneros *Arctocephalus* y *Zalophus*.

Esta información sería relevante para la planificación de investigaciones de campo sobre las especies de lobos marinos de estos géneros.

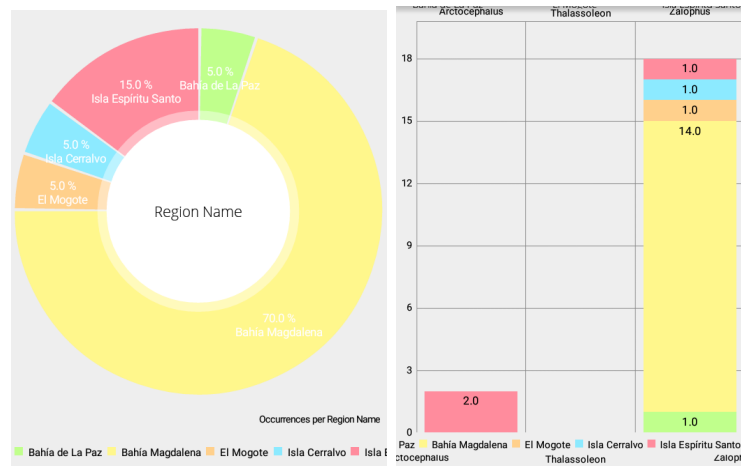


Figura 5.5: Captura de la interfaz de la aplicación móvil, que muestra las ocurrencias de lobos marinos (familia Otariidae) en las distintas regiones geográficas, diferenciando los géneros.

### 5.2.5. ¿Cómo ha variado la forma de registrar la biodiversidad de la región a los largo del tiempo?

Para contestar esta pregunta, utilizamos las dimensiones “Basis of record” y “Time”, para mostrar sus métricas dentro de los renglones y columnas de las tablas y gráficos resultantes. El producto de esta consulta multidimensional se puede observar en la fig. 5.6.

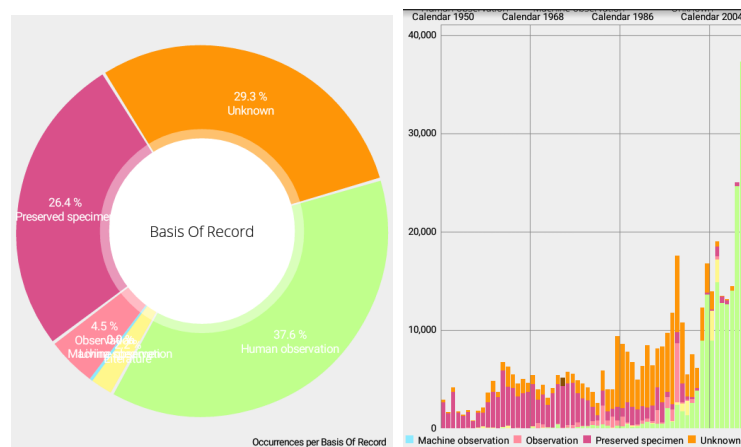


Figura 5.6: Captura de la interfaz de la aplicación móvil, que muestra los tipos de registros que se han realizado en un periodo de tiempo, que va desde 1950 hasta 2013.

Como se puede observar, ha ocurrido un cambio en la tendencia de los tipos de registros de la diversidad biológica de la región ya que, a partir del final del siglo XX, la observación humana



(Human observation) es la tendencia que ha acabado por predominar, dejando el registro a partir de especímenes preservados (Preserved specimen) para los registros históricos. Esto es un dato importante, ya que nos habla de un cambio en la forma de llevar a cabo estudios sobre la biodiversidad de la región, y debería ser tomado en cuenta para los trabajos futuros en los campos de biología y ecología que aquí ocurran.

Y es de esta forma que, combinando distintas dimensiones y jerarquías del modelo podemos dar respuesta a gran diversidad de preguntas relevantes sobre la variedad de organismos encontradas en la región noroeste del territorio mexicano.

# Capítulo 6

## Conclusiones y trabajo futuro

Se logró comprobar el potencial de las aplicaciones móviles para utilizarse como herramientas para la investigación en campo, gracias a sus funciones de geoposicionamiento geográfico y sus capacidades de almacenamiento interno. Así, con la implementación final de este sistema, el dispositivo móvil funciona como una fuente de información primaria para el estudio de la diversidad de seres vivos en una región particular, al mismo tiempo que actúa como una herramienta para el análisis de información histórica desde múltiples dimensiones, sintetizándola y adaptándola a la interfaz del móvil para facilitar su comprensión.

Con este desarrollo se muestra además una forma en la que las aplicaciones móviles pueden ser de gran ayuda en el terreno de la investigación científica, al integrarse con las tecnologías de inteligencia de negocios, como las bodegas de datos y los hipercubos, de modo que sus limitaciones de espacio de almacenamiento y potencia de procesamiento pueden ser superadas al pasar la mayor carga de trabajo al servidor y así aprovechar sus capacidades de movilidad, para llegar a lugares donde las computadoras tradicionales no están disponibles. Esto cobra una especial relevancia al encontrarnos en un contexto en el que los dispositivos móviles están en pleno auge, ya que abre nuevas posibilidades para la innovación en el desarrollo de aplicaciones para la investigación.

Con la implementación de este modelo de datos se comprobó también el potencial de los modelos multidimensionales para analizar la información histórica sobre la biodiversidad, permitiendo

combinar distintas perspectivas para lograr dar respuesta a preguntas complejas, como las revisadas en el capítulo de Resultados. Gracias a esta herramienta de análisis es posible brindar información relevante para la investigación en las áreas de biología y ecología, dando además soporte para la toma de decisiones públicas relacionadas con la biodiversidad, como el desarrollo de políticas para el uso de la tierra y la designación de áreas naturales protegidas.

De esta forma, se aportó una nueva herramienta completamente funcional que facilita el análisis de la información sobre biodiversidad en la región, para la investigación y la toma de decisiones, al mismo tiempo que brinda una importante ayuda en el trabajo de campo para los investigadores y estudiantes.

Por otro lado, este modelo multidimensional se diseñó de forma que puede ser extrapolado para cubrir un territorio mucho más grande, de modo que pueda analizarse una base de datos de biodiversidad extraída de la GBIF que abarque el territorio mexicano completo, incluyendo su zona marítima. Por lo tanto, sería posible realizar un análisis mucho más profundo de la distribución de especies en las diferentes regiones de nuestro país y contar con la información desde una aplicación móvil, durante el trabajo de campo que se realice en las mismas.

Para lograr este trabajo a mayor escala, sería necesario considerar los requerimientos técnicos de esta nueva implementación, los cuales incluyen:

- La capacidad de almacenamiento físico, al tratarse de una base de datos intermedia que contendría más de un millón y medio de registros.
- La capacidad de procesamiento del servidor que contenga la nueva base de datos y el hipercubo, los cuales crecerían en tamaño y tiempo de ejecución de las consultas.
- Los requerimientos de optimización particular en el proceso de creación de las regiones de interés, ya que en el modelo actual tiene un alto costo de tiempo de procesamiento, lo cual se vería incrementado al aumentar notablemente la cantidad de registros. Para solventar este problema se podrían estudiar soluciones basadas en el desarrollo de cómputo paralelo, como el multiproceso, los clústeres de computadoras o el multiprocesamiento con GPU.
- Los requerimientos en licencias de software al tratarse ya no solo de un trabajo académico

sino de un sistema para uso permanente, lo cual podría llevarnos a necesitar reemplazar los componentes de software propietario por software libre.

Todas estas tareas futuras, aunque requerirían de un mayor esfuerzo en su desarrollo, tendrían la gran ventaja de crear nichos de oportunidad para más profesionales y estudiantes de posgrado, los cuales podrían seguir desarrollando trabajos académicos y proyectos de investigación, dentro del área de la informática sobre biodiversidad.

# Apéndice A

## Instrucciones SQL para la inserción de las regiones

```
INSERT INTO region (name) VALUES('Isla Espiritu Santo');
INSERT INTO vertex (decimalLongitude, decimalLatitude, regionId,
vertexNumber)
VALUES(-110.408316, 24.619429, 2, 1),
(-110.333472, 24.550121, 2, 2),
(-110.258628, 24.447024, 2, 3),
(-110.317679, 24.392005, 2, 4),
(-110.365058, 24.387627, 2, 5),
(-110.419989, 24.474525, 2, 6),
(-110.429602, 24.576350, 2, 7);
```

```
INSERT INTO region (name) VALUES('Bahia Magdalena');
INSERT INTO vertex (decimalLongitude, decimalLatitude, regionId,
vertexNumber)
VALUES(-112.135442, 25.274719, 3, 1),
(-112.073644, 25.264784, 3, 2),
(-112.054418, 24.887899, 3, 3),
```

```
(-111.774693, 24.611868, 3, 4),
(-111.811772, 24.524441, 3, 5),
(-112.001286, 24.506948, 3, 6),
(-112.166081, 24.648070, 3, 7),
(-112.307530, 24.794020, 3, 8);
```

```
INSERT INTO region (name) VALUES('Isla Cerralvo');
INSERT INTO vertex (decimalLongitude, decimalLatitude, regionId,
vertexNumber)
VALUES(-109.931940, 24.396153, 4, 1),
(-109.769892, 24.158302, 4, 2),
(-109.789118, 24.117573, 4, 3),
(-109.862589, 24.130106, 4, 4),
(-109.941553, 24.271024, 4, 5),
(-109.953913, 24.327973, 4, 6);
```

```
INSERT INTO region (name) VALUES('Sierra de la Laguna');
INSERT INTO vertex (decimalLongitude, decimalLatitude, regionId,
vertexNumber)
VALUES(-109.920178, 23.796801, 5, 1),
(-109.769116, 23.701269, 5, 2),
(-109.775983, 23.187223, 5, 3),
(-109.920178, 23.095039, 5, 4),
(-110.023175, 23.182173, 5, 5),
(-110.030041, 23.710071, 5, 6);
```

```
INSERT INTO region (name) VALUES('Bahia de La Paz');
INSERT INTO vertex (decimalLongitude, decimalLatitude, regionId,
vertexNumber)
VALUES(-110.641062, 24.317420, 6, 1),
```

(-110.334132, 24.318671, 6, 2),  
(-110.317652, 24.245441, 6, 3),  
(-110.298769, 24.241371, 6, 4),  
(-110.308039, 24.220708, 6, 5),  
(-110.296550, 24.218656, 6, 6),  
(-110.297587, 24.172486, 6, 7),  
(-110.343248, 24.137401, 6, 8),  
(-110.342562, 24.114528, 6, 9),  
(-110.403673, 24.098859, 6, 10),  
(-110.427706, 24.120481, 6, 11),  
(-110.433886, 24.178437, 6, 12),  
(-110.564005, 24.209441, 6, 13),  
(-110.621201, 24.260118, 6, 14);

# Bibliografía

Abd El-Aziz, A. and Kannan, A. JSON encryption. In *2014 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6, Jan. 2014. doi: 10.1109/ICCCI.2014.6921719.

Aguilar Yelpiz, E. *Modelo de datos multidimensional en el ámbito prehospitalario de Cruz Roja Mexicana delegación La Paz, Baja California Sur*. Tesis de maestría, Instituto Tecnológico de La Paz, 2012.

Balke, M., Schmidt, S., Hausmann, A., Toussaint, E. F., Bergsten, J., Buffington, M., Häuser, C. L., Kroupa, A., Hagedorn, G., Riedel, A., Polaszek, A., Ubaidillah, R., Krogmann, L., Zwick, A., Fikáček, M., Hájek, J., Michat, M. C., Dietrich, C., Salle, J. L., and Mantle, B. Biodiversity into your hands - a call for a virtual global natural history 'metacollection'. *Frontiers in Zoology*, 10(1):1–9, Oct. 2013. ISSN 17429994. doi: 10.1186/1742-9994-10-55. URL <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=91262053&site=ehost-live>.

Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G. Falling Asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application Usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 47–56, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0541-9. doi: 10.1145/2037373.2037383. URL <http://doi.acm.org/10.1145/2037373.2037383>.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms*. The MIT Press, 2009.



- Dana, P. H. Geodetic datum overview. *The Geographer's Craft Project*, 1995. URL <http://www.colorado.edu/geography/gcraft/notes/datum/datum.html>.
- Date, C. *Introducción a los sistemas de bases de datos*. Pearson Educación, 2001.
- Do, Q., Martini, B., and Choo, K.-K. R. Exfiltrating data from android devices. *Computers & Security*, 48:74–91, Feb. 2015. ISSN 0167-4048. doi: 10.1016/j.cose.2014.10.016. URL <http://www.sciencedirect.com/science/article/pii/S016740481400162X>.
- Edwards, J. L., Lane, M. A., and Nielsen, E. S. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, 289(5488):2312–2314, Sept. 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.289.5488.2312. URL <http://www.sciencemag.org.etechnology.idm.oclc.org/content/289/5488/2312>.
- Guralnick, R. and Hill, A. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics*, 25(4):421–428, Feb. 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btn659. URL <http://bioinformatics.oxfordjournals.org/content/25/4/421>.
- Hardisty, A., Roberts, D., and Community”, T. B. I. A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, 13(1):16, Apr. 2013. ISSN 1472-6785. doi: 10.1186/1472-6785-13-16. URL <http://www.biomedcentral.com/1472-6785/13/16/abstract>.
- Hill, A. W., Guralnick, R., Flemons, P., Beaman, R., Wieczorek, J., Ranipeta, A., Chavan, V., and Remsen, D. Location, location, location: utilizing pipelines and services to more effectively georeference the world’s biodiversity data. *BMC Bioinformatics*, 10:1–9, Jan. 2009. ISSN 14712105. URL <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=47327572&site=ehost-live>.
- Hill, L. L. *Georeferencing: The Geographic Associations of Information*. MIT Press, 2009. ISBN 9780262512527.
- Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P. *Fundamentals of Data Warehouses*. Springer, 2002.

- Lane, M. A. L., James, L. E., Los, W., H. J. Hof, C., Berendsohn, W. G., Geoffroy, M., Scoble, M. J., MacLeod, N., O'Neill, M., Walsh, S. A., Curry, G. B., Connor, R. J., Jones, A. C., Triebel, D., Persoh, D., Nash, T. H., Zedda, L., Rambold, G., White, R. J., Sterling, J. A., Seberg, O., Humphries, C. J., Borchsenius, F., and Dransfield, J. *Biodiversity Databases*. Systematics Association Special Volumes. CRC Press, first edition, Apr. 2007. URL <http://www.crcpress.com/product/isbn/9780415332903>.
- Núñez, I., González-Gaudio, E., and Barahona, A. La biodiversidad: historia y contexto de un concepto. *Interciencia*, 28(7):387–393, 2003. URL [http://scielo.org.ve/scielo.php?pid=s0378-18442003000700006&script=sci\\_arttext](http://scielo.org.ve/scielo.php?pid=s0378-18442003000700006&script=sci_arttext).
- Olaya, V. *Sistemas de Información Geográfica*. Capítulo hispano-hablante de OSGeo, 2014. URL <http://volaya.github.io/libro-sig/>.
- Otegui, J., Ariño, A. H., Encinas, M. A., and Pando, F. Assessing the primary data hosted by the spanish node of the global biodiversity information facility (GBIF). *PLoS ONE*, 8(1): 1–15, Jan. 2013. ISSN 19326203. doi: 10.1371/journal.pone.0055144. URL <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=85384211&site=ehost-live>.
- Rands, M. R. W., Adams, W. M., Bennun, L., Butchart, S. H. M., Clements, A., Coomes, D., Entwistle, A., Hodge, I., Kapos, V., Scharlemann, J. P. W., Sutherland, W. J., and Vira, B. Biodiversity conservation: Challenges beyond 2010. *Science*, 329(5997):1298–1303, Oct. 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1189138. URL <http://www.sciencemag.org.etechnology.idm.oclc.org/content/329/5997/1298>.
- Rucker, M. Encyclopedia of life. *Reference Reviews*, 28(1):29–30, Jan. 2014. ISSN 0950-4125. doi: 10.1108/RR-09-2013-0238. URL <http://www.emeraldinsight.com.etechnology.idm.oclc.org/doi/full/10.1108/RR-09-2013-0238>.
- Schnase, J. L., Cushing, J., and Smith, J. A. Biodiversity and ecosystem informatics. *Journal of Intelligent Information Systems*, 29(1):1–6, Aug. 2007. ISSN 0925-9902, 1573-7675. doi: 10.1007/s10844-006-0027-7. URL <http://link.springer.com/article/10.1007/s10844-006-0027-7>.
- Vaisman, A. and Zimányi, E. *Data Warehouse Systems*. Springer, 2014.

Vihervaara, P., Ronka, M., and Walls, M. Trends in ecosystem service research: Early steps and current drivers. *Ambio*, 39(4):314–324, June 2010. ISSN 0044-7447. doi: 10.1007/s13280-010-0048-x. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3357705/>.

Wieczorek, J., Guo, Q., and Hijmans, R. J. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745–767, Dec. 2004. ISSN 13658816. doi: 10.1080/13658810412331280211. URL <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=15374169&site=ehost-live>.