UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Biocuración de literatura

SEMINARIO DE TITULACIÓN

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGO

P R E S E N T A :

Roberto Santos Solórzano

DIRECTORA DE SEMINARIO:

Dra. Layla Michán Aguirre

2015

Ciudad Universitaria, D. F.

Hoja de datos de jurado

1. Datos del alumno
Santos
Solórzano
Roberto
56 45 54 75
Universidad Nacional Autónoma de México
Facultad de Ciencias
Biología
099314868

2. Datos del tutor
Dra.
Layla
Michán
Aguirre

3. Datos del sinodal 1
Dr.
Frank Raúl
Gío
Argáez

4. Datos del sinodal 2
M. en C.
Patricia
Magaña
Rueda

5. Datos del sinodal 3
M. en I.
Adrián
Girard
Islas

6. Datos del sinodal 4
Lic. En I.
Beatriz Adriana
González
Alvarado

7. Datos del trabajo escrito
Biocuración de literatura
47 p
2015

Estoy satisfecho con el misterio de la eternidad de la vida y con el conocimiento, el sentido, de la maravillosa estructura de la existencia. Con el humilde intento de comprender aunque más no sea una porción diminuta de la Razón que se manifiesta en la naturaleza.

Albert Einstein

A mis Padres, María del Carmen y José, porque con su amor, tenacidad y ejemplo, han forjado en mí al hombre que soy, que me enseñaron que la calidad humana se encuentra en los valores que nos preceden, que la inteligencia no está en los títulos o grados que ostentamos y que el éxito sólo es la conclusión del trabajo, esfuerzo y constancia que hayamos puesto.

AGRADECIMIENTOS

Agradecimientos académicos:

A la Universidad Nacional Autónoma de México y a la Facultad de Ciencias por ser mi segundo hogar.

A la Dra. Layla Michán Aguirre y al Laboratorio de Cienciometría, Información e Informática Biológica (CIIB), en especial a la Biol. Diana Ramírez Álvarez coordinadora del Sistema de Información Ciencias, UNAM. Además agradezco al IIIC, AC por el apoyo en cómputo. José Ángel Bautista Ruiz, Fernando Flores Toledo y Octavio Aldebarán Sandoval Maldonado especialistas en Sistemas y Cómputo del CIIB, FC UNAM. Así como a Mario Arturo Pérez Rangel y Beatriz Adriana González Alvarado Técnicos Académicos del Departamento de Cómputo de la Facultad de Ciencias, UNAM por su asesoría para los servidores y redes.

A mis sinodales Dr. Frank Raúl Gío Argáez, M. en C. Patricia Magaña Rueda, M. en I. Adrián Girard Islas y Lic. En I. Beatriz Adriana González Alvarado no solo por sus comentarios y sugerencias que enriquecieron mi trabajo, sino por la sensibilidad y apoyo mostrado para concluir éste proceso.

Esta investigación se llevó a cabo gracias al financiamiento de CONACYT, Ciencia Básica. Proyecto 13276 “Análisis de las ciencias biológicas en la actualidad 1980-2010” y DGAPA, PAPIME, Proyecto PE212112 “Web 2.0 y 3.0 para dominio de la literatura biológica”.

Agradecimientos personales:

A la Dra. Layla Michán, por ser mi directora de tesis, por permitirme desarrollar mi proyecto en su laboratorio, por enseñarme una nueva visión y dimensión de la biológica, por hacerme crecer profesionalmente y brindarme la oportunidad de colaboración.

A mi Madre, María del Carmen Solórzano, porque con tu Amor, Fortaleza, Dedicación y Sacrificio siempre nos sacaste adelante, eres el motor de mi espíritu y mi inspiración.

¡Conseguimos unas de las metas mamá, aún nos faltan muchas más!

A mi Padre, José Santos, porque con tu Cariño, Gallardía, Solidez y Empeño me mostraste que es ser un “Hombre” y voy en camino a serlo. Porque es un gran orgullo ser tu hijo, porque eres mi ejemplo, porque sin tu ayuda y presencia esto no hubiera sido posible.

A mis Hermanos, Arturo y Elizabeth, porque con su Cariño, Confianza y Apoyo incondicional siempre me han impulsado a ser mejor, porque tenemos virtudes distintas y las hemos complementado e implementado en nuestra vida. Siempre cuentan conmigo.

Al amor de mi vida, Erika “Rik” porque con tu Amor, Paciencia, Confianza y Esfuerzo me mostraste otro mundo en el que hemos vivido por tantos años. Siempre estarás en mi corazón.

A Edith, por tu enorme Cariño, Entrega, Lealtad y Solidaridad, porque sin tu ayuda nunca hubiera alcanzado esta meta.

A mi mejor amigo, Cesar, porque siempre creíste en mí, por incluirme en tu vida, porque ahora somos familia.

A mi carnal, Reinchard, por tu confianza, porque creamos algo mucho más grande que sólo una amistad.

A mis súper cuates, Paco, Jon y Tavo, por las incontables ocasiones que me han demostrado ser verdaderos amigos.

A los Bullets, Sig, Pavel y José, por su influencia siempre positiva.

A los Montañeses, Alvaro, Moy e Isra, por las tantas horas de amenidad.

A todo el Laboratorio de Cienciometría, Información e Informática Biológica (CIIB), Facultad de Ciencias, UNAM.

En especial a Diana por siempre brindarme tu apoyo y por alentarme a ser mejor.

A Daniel por coordinar mi trabajo, tus comentarios y sugerencias enriquecieron mi conocimiento, mi trabajo y el gusto por esto que hacemos.

A Rox y Ale, porque forjamos una amistad, por las muchas horas de almuerzo, chismes y amenidad que hicieron mi estancia más placentera y que no me dejaron claudicar.

A Tania, porque con tu amistad y alegría los días fueron más alivianados.

A Vane, Olga, Eniak y David, por sus consejos, por las charlas, y por ser un gran equipo de trabajo.

Índice

INTRODUCCIÓN 1

Información científica en la actualidad 1
Desventajas del formato impreso, ventajas del formato digital y la cultura de datos 3
Bases de datos 13
Biocuración 15

ANTECEDENTES 20

OBJETIVOS 24

General: 24
Particulares: 24

EXPOSICIÓN DEL TEMA 25

CONCLUSIONES 42

REFERENCIAS 44

INTRODUCCIÓN

Las colecciones biológicas de interés taxonómico, como los museos, herbarios, zoológicos, jardines botánicos y bibliotecas, entre otras, desempeñan un papel especial dentro de la biología debido a que uno de sus propósitos es conservar a largo plazo los varios millones de materiales (objetos) y la información asociada a cada uno de ellos, producto de la practica biológica realizada durante más de tres siglos (Thessen y Patterson, 2011). Sin embargo, debido a que en las últimas décadas se ha presentado una producción exponencial en la cantidad y diversidad de datos e información relacionada con la biología, se han hecho necesarias medidas revolucionarias para su gestión (Howe *et al.*, 2008).

Por ello, en la actualidad han aparecido nuevos conceptos para describir este fenómeno, como por ejemplo: “Big Data” que no sólo se refiere o está representado, por el aumento colosal del volumen y detalle de la información biológica, sino también por el desarrollo de la *Web,* las herramientas, aplicaciones, programas, etc., disponibles para la práctica de los científicos (Manyika *et al.*, 2011).

Información científica

Simultáneamente al crecimiento exponencial en la cantidad y diversidad de datos e información producida mediante las investigaciones biológicas; durante la segunda mitad del siglo XX se presentó también un amplio desarrollo en las tecnologías de la información y la comunicación (TIC’s) por lo que se adoptó el uso de las computadoras como herramientas indispensables en la práctica cotidiana de los biólogos, generando, una (r)evolución informática, con lo que se ha aumentado considerablemente la capacidad para generar, sistematizar, compartir, transmitir, analizar y difundir la información, con influencia directa en la investigación científica (Llorente-Bousquets y Michán, 2010; Michán, 2011)

En este nuevo siglo en la biología, al igual que en otras áreas, se ha hecho énfasis en el uso adecuado de la información, pero sobre todo, en el conocimiento generado a partir de los valores agregados a ella. En este contexto, se ha empezado a hablar de la gestión del conocimiento, pero más bien, debiera tratarse, como la gestión de la información para la generación de nuevos acervos que influirán en el desarrollo de productos de alto valor agregado, ya que, el conocimiento sólo puede ser gestionado cuando éste se traduce como información, que además es indispensable que esta información esté reunida, procesada, organizada, almacenada y diseminada, por ejemplo, mediante

bases de datos, redes de información compartida, comunidades virtuales, entre otros medios de transferencia de datos e información (Morales *et al.*, 2004). Por todo esto el referirse a información biológica el día de hoy implica como menciona Michán, 2011, en su artículo: "*Cienciometría, información e informática en ciencias biológicas: Enfoque interdisciplinario para estudiar interdisciplinas"* que tengamos que conocer, entender y desarrollar: términos, métodos y teorías, como: sociedad del conocimiento, sociedad de la información, globalización, infodiversidad, acceso abierto, e-ciencia, e-investigación, grid, colaboratorios, conocimiento basado en la literatura, minería de textos, *Web* semántica, índice de impacto, cocitación, *Web 2.0* y 3.0, redes sociales, plagio, recuperación de información, democratización, cómputo en nube, derecho al olvido, semántica, solo por mencionar algunas de ellas.

Cabe mencionar que esta evolución no sólo ha cambiado la manera en cómo se genera el conocimiento, información y datos, sino también ha revolucionado la forma en que se presenta, transmite y difunde, por ejemplo, los documentos científicos (cuadro 1), pueden ser publicados de forma impresa o electrónica, y su información no sólo consiste en texto, sino también pueden ser imágenes, sonidos (audio) y/o video, por nombrar algunas. Otras de las medidas adoptadas en la actualidad son la forma en cómo se procesa, ordena, sistematiza y analiza la información, como ya se dijo en el párrafo anterior, entre las más comunes se encuentra el uso de bases de datos. A pesar de haber muchas variedades de bases de datos, en este trabajo nos centraremos en las bases de datos bibliográficas digitales, que de manera general pueden ser definidas como programas que permiten guardar, ordenar, procesar y presentar los datos relacionados con los documentos o textos de interés científico, y su contenido o acceso está disponible en la *Web* de manera libre o restringida; a su vez estas bases de datos forman parte de grandes colecciones bibliográficas digitales, por ejemplo: 1) e-bibliotecas, 2) sistemas de información, 3) catálogos e índices, 4) editoriales y revistas y/o 5) repositorios (Castellanos Morales, 2013). De estas cinco en este trabajo sólo se mencionarán los repositorios y los sistemas de información, ya que la base de datos TaXMeXX, que es el centro de este proyecto y que se abordará en detalle más adelante, pertenece al Sistema de Información Ciencias, UNAM, y al Repositorio de la Facultad de Ciencias, UNAM.

Cuadro 1: Tipos de documentos científicos

1. Libros	**2. Artículos (revistas)**
3. Series	**4. Memorias**
5. Tesis	***6. Pre prints***
7. Reseñas	**8. Ensayos**
9. Monografías	**10. Listas y catálogos**
11. Bibliografías	**12. Currículos**
13. Películas	**14. Fotografías**
15. Materiales electrónicos	**16. Otros**

Tomado y modificado de: Michán, L. y Llorente, J. 2003. *La taxonomía en México durante el siglo XX.* Publicaciones especiales del museo de zoología, Número 12. Facultad de Ciencias, UNAM. México. 237 pp.

Nota: Esta clasificación está hecha para los fines prácticos de este proyecto con base en la metodología de la base de datos TaXMeXX.

La evolución de la información durante la transición del siglo XX al XXI ha sido drástica, ya que este fenómeno ha repercutido en todos los ámbitos de la práctica biológica, tanto en la forma de producir conocimiento científico como en el desarrollo de nuevos campos del conocimiento (Michán *et al*, 2010), por ejemplo, la bioinformática. También se ha transformado la práctica científica a varios niveles, desde el objeto de estudio hasta el costo y el tiempo requeridos para la generación, procesamiento y el análisis de la información, las técnicas, las tecnologías, herramientas utilizadas e incluso la comunicación y colaboración entre los científicos, especialistas, alumnos, grupos, instituciones, organizaciones, etc., involucradas con la biología hoy en día (Michán, 2011).

Desventajas del formato impreso, ventajas del formato digital y la cultura de datos

Como se puede ver en el cuadro 1, existen muchos tipos de documentos científicos, la mayoría de ellos contienen texto e imágenes y prácticamente, durante todo el siglo pasado, estos fueron presentados en formato impreso siguiendo un modelo establecido mucho antes de la era de las computadoras, el espacio barato de almacenamiento electrónico, y la publicación digital.

Este formato trae consigo ciertas desventajas y limita la comunicación científica (Rzhetsky *et al.*, 2008) ya que:

- Los científicos no pueden registrar y compartir todas sus conclusiones.
- Algunos hechos son considerados demasiado triviales para incluirse en el texto.
- Hallazgos aislados o resultados negativos son a menudo retenidos del registro publicado.
- Algunos conjuntos de datos son demasiado grandes para incluirse.
- Los costos de impresión así como de inscripción suelen ser altos.
- Por otro lado, la divulgación y difusión de la revista depende de la región geográfica, las instituciones relacionadas, las leyes y reglamentaciones locales e internacionales, etc., por lo que sólo tendrán acceso a ella un número limitado de personas.
- Otro problema es lo limitado de la extracción de información y datos, el almacenamiento a largo plazo, el análisis y procesamiento de un gran volumen de publicaciones, las búsquedas de información, etc.
- Se requiere de mucho tiempo para publicar los trabajos.

Durante la segunda mitad del siglo XX, han surgido decenas de miles de revistas que tratan sobre temas biológicos, y la avalancha de nuevos artículos en las ciencias está dando lugar a una sobrecarga de información (Rzhetsky *et al.*, 2008) por ello, en la actualidad, los avances científicos y tecnológicos demandan nuevos y mejores métodos para el tratamiento de toda esta información en beneficio de necesidades específicas (Castro *et al.*, 2005). La aceptación de la publicación electrónica y la adopción del formato digital está brindando una alternativa para enfrentar de mejor manera estas problemáticas (Michán, 2011), dado que han resultado ser un vehículo eficiente de diseminación de contenidos. También resulta ser uno de los principales medios de comunicación de las comunidades científicas, facilitando el intercambio desde distintos puntos geográficos (Castro *et al.*, 2005), sin embargo, es importante mencionar que a pesar de lo evidente de las muchas ventajas (Cuadro 2), también traen consigo algunas desventajas (Cuadro 3).

Cuadro 2: Ventajas de las publicaciones electrónicas.

1. Preservan documentos raros y frágiles, sin limitar el acceso a quienes deseen consultarlos.
2. Facilitan la transmisión mediante redes informáticas.
3. Posibilitan el acceso a muchos usuarios simultáneamente
4. Ofrecen solución al problema de espacio físico para el almacenamiento.
5. Disminuyen costos de edición y de distribución al utilizar los medios electrónicos para la transmisión de la información.
6. Permiten búsquedas más ágiles en el texto completo.
7. Facilitan el acceso instantáneo sin necesidad de desplazamiento.
8. Proporciona enlace a otros recursos relacionados, como películas y animaciones, que facilitan la expresión de ideas difíciles de plasmar en un formato impreso.
9. Establecen una relación más cercana entre autores y lectores, por correo electrónico, favoreciendo la comunicación científica.
10. Permiten la publicación inmediatamente, a partir de un régimen de edición continua.
11. Posibilitan incorporar correcciones y comentarios hechos por los lectores.
12. Disminuyen costos en el consumo de papel, ya que se hacen copias impresas sólo de los artículos que realmente son de interés.
13. Se puede guardar la colección completa de uno o varios títulos.
14. Eliminan muchos de los pasos de la publicación impresa en la relación entre editores/proveedores y la biblioteca: costos de impresión, encuadernación, empaque, distribución, transporte, tarifa postal y almacenamiento.

Información tomada y modificada de: Castro, A; Olivares, S. S; Alonso, J. O; Ramírez, M. E: Algunas reflexiones sobre la revista electrónica en la UNAM. *Revista Digital Universitaria*. México. 2005. Volumen 6. Número 4. ISSN: 1067-6079. Publicado en línea.http://www.revista.unam.mx/vol.6/num4/art37/abr_art37.pdf

Cuadro 3: Desventajas de las publicaciones electrónicas.

1. Inversión inicial considerable, aunque a largo plazo resultan más económicas.
2. Altos costos de suscripción a la publicación
3. Incomodidad de visualización en pantalla, aunque el formato PDF es un avance en este sentido.
4. Barrera idiomática en nuestros países toda vez que los recursos son en su mayoría en idioma inglés.
5. Se requiere de conexión a Internet con costos adicionales relativos a la infraestructura de telecomunicaciones.
6. Desconfianza de los científicos con respecto a los derechos de autor.
7. La mayoría de los documentos se encuentran en acceso restringido, ya que son propiedad de las editoriales.
8. En acceso libre se encuentran principalmente *preprints* y muchas veces no se cuentan con las versiones finales.
9. Uso de tecnología que se puede convertir en obsoleta en poco tiempo.
10. Garantizar el acceso futuro a los contenidos de la revista, lo cual se ve en peligro si a los archivos se les cambia el nombre o la ubicación.
11. Podrían desaparecer de Internet si se cierra la institución o empresa que las publica.

Información tomada y modificada de: Castro, A; Olivares, S. S; Alonso, J. O; Ramírez, M. E: Algunas reflexiones sobre la revista electrónica en la UNAM. *Revista Digital Universitaria.* México. 2005. Volumen 6. Número 4. ISSN: 1067-6079. Publicado en línea. http://www.revista.unam.mx/vol.6/num4/art37/abr_art37.pdf

Sin embargo, a pesar de las desventajas mencionadas las publicaciones electrónicas disponibles a través de la *Web* nos ofrecen: la personalización, automatización, actualización e inmediatez de la información (Michán, 2011). Hoy en día muchos de los documentos científicos impresos se están transformando en documentos electrónicos para su tratamiento, transmisión, almacenamiento, recepción, etc., por medio de ordenadores, por lo que se exige que estos documentos estén en un

formato compatible con el propio ordenador, a este proceso se le conoce como digitalizar o digitalización. De este modo apareció el concepto de **literatura digital** que se remonta a la década de los 80's y que de manera general podríamos definir como todas las obras en formato digital (Wikipedia, 2014), y que reiterando lo dicho antes, presenta las ventajas, de que con el desarrollo tecnológico y la creación de nuevas y mejores herramientas se facilita la obtención de información y de datos, que pueden ser almacenados, procesados, analizados, difundidos, divulgados, compartidos, consultados, actualizados, manejados, entre otras muchas acciones que mejoran las investigaciones y estudios. Sin embargo, existen diversas cuestiones técnicas y sociológicas que enfrenta la biología a medida que se transforma en una disciplina más centrada en los datos. Tres desafíos principales son: 1) la falta de normas generales; 2) la falta de incentivos a los científicos para compartir datos; 3) la falta de infraestructura y apoyo adecuados (Thessen y Patterson, 2011).

A pesar de que los datos son considerados como la base de la información, el término "datos" no se utiliza de manera consistente (Thessen y Patterson, 2011), ya que un dato puede ser: 1) un antecedente necesario para llegar al conocimiento exacto de algo o para deducir las consecuencias legítimas de un hecho, 2) un documento, testimonio, fundamento o bien 3) Información dispuesta de manera adecuada para su tratamiento por un ordenador (RAE, 2014) lo que es importante mencionar es que los datos son, en gran medida, independientes de contexto, análisis u observador. Como los datos se convierten en limitados, se filtran y se seleccionan, adquieren o se les asigna un significado en el contexto donde se aplican. Esto es parte del proceso que transforma los datos en información (Thessen y Patterson, 2011; Sánchez *et al.*, 2011).

Para gestionar el conocimiento se requiere que sea traducido en información, y que además esta información sea a su vez diseminada en datos. Estos datos pueden ser usados posteriormente por investigadores, con lo que se genera más información y por ende nuevo conocimiento (Sánchez *et al.*, 2011) (figura 1). La necesidad urgente de comprender los fenómenos globales complejos, el diluvio de datos derivados de las nuevas tecnologías y la mejora de la gestión de datos están impulsando una agenda para extender a la biología con más dimensiones de descubrimiento basados en datos, esto es, la comprobación de hipótesis y el descubrimiento de conocimientos científicos a través de la gestión y análisis de datos pre-existentes. El descubrimiento basado en datos se basa en el acceso y la

reutilización de los datos que muy probablemente se han generado para hacer frente a otros problemas científicos, el descubrimiento impulsado por los datos (modelos *in silico*) contrasta con el proceso más familiar de la investigación científica basada en la recopilación de nuevos datos, ya sea por la experimentación o al hacer nuevas observaciones en modelos *in vivo* o *in vitro*, requiere un gran sistema abierto de datos a través de toda la amplitud de la biología y hacia las disciplinas adyacentes (Michán, 2011; Thessen y Patterson, 2011).

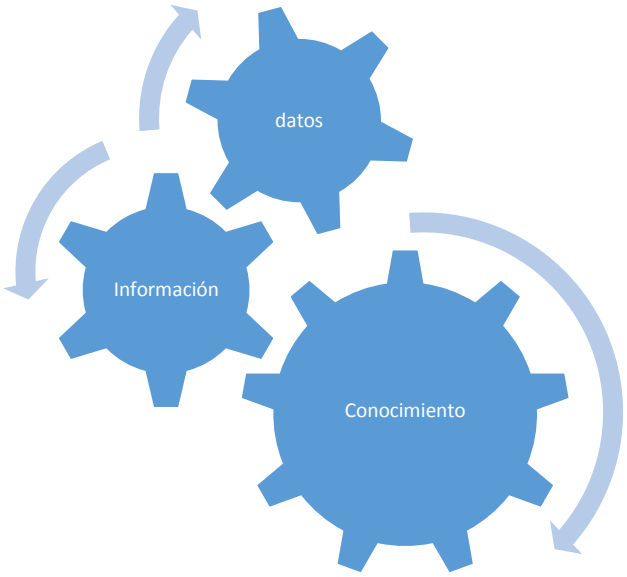


Figura 1: El conocimiento genera información y esta se disemina en datos. Entonces estos datos pueden utilizarse para generar nueva información y por ende nuevo conocimiento.

A la par que las computadoras personales y el acceso a Internet se han convertido en parte integral de la investigación biológica, los biólogos están publicando cada vez más datos a través de repositorios, sus propios sitios *Web*, o están participando en los entornos de colaboración como los que permite que los datos se capturen mediante servicios ofrecidos por diversas páginas o plataformas en la red, como por ejemplo, eBird (www: //ebird.org). Sin embargo, para que estos cambios dominen en toda la amplitud de la disciplina e influyan en el ciclo de vida completo de los datos es necesario tener una cultura de datos, esta frase se refiere a la utilización de los datos explícitos e implícitos y a las expectativas que determinan el destino de los datos.

La **cultura de datos** se refiere a las convenciones sociales de la adquisición, conservación, preservación, uso compartido y la reutilización de los datos. Su objetivo es hacer que sean: digitales, estandarizados y abiertamente accesibles en un formato reutilizable.

La preparación de los datos para su reutilización a menudo implica una serie de pasos o etapas que se relacionan con la captura, digitalización, estructura, almacenamiento, conservación, acceso y movilidad, una rica diversidad de herramientas, servicios y aplicaciones para analizarlos y visualizarlos (Thessen y Patterson, 2011) (Figura 2).

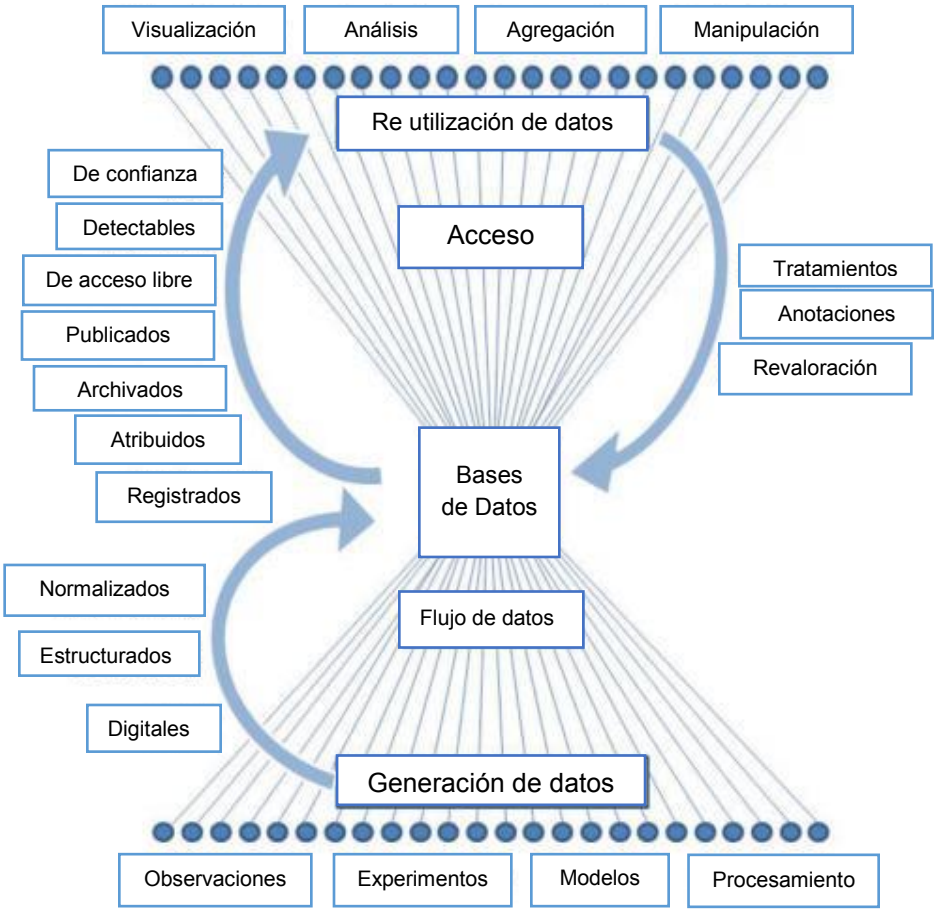


Figura 2: La cultura de datos y la reutilización de los mismos, para ello se requiere de varios procesos aquí ilustrados y descritos más adelante.

Imagen tomada y modificada de: Thessen y Patterson, 2011.Data issues in the life sciences. *Zookeys. 2011; (150): 15-51. Publicado en linea Nov 28, 2011. Doi: 10.3897/zookeys.150.1766*

Para cumplir con el objetivo de la cultura de datos se requiere que:

1. **Los datos sean retenidos-** Relativamente pocos datos adquiridos históricamente se han mantenido en una forma accesible por los científicos, proyectos o instituciones. La cultura de la eliminación de los datos después de la publicación, la terminación de una concesión, la reubicación o retiro de un científico son algunos de los casos por medio de los cuales se pueden perder datos. Existe todavía un debate sobre si se deben conservar todos los datos, o si los subconjuntos de datos deben ser seleccionados para la retención, o si los datos conservados deben estar sujetos a una revisión periódica.
2. **Los datos sean digitales.** Esto facilita su análisis, accesibilidad, movilidad y conservación, además, permite abordar de muy diferentes maneras la información.
3. **Los datos deben estructurarse-** Los datos digitales pueden ser no estructurados (por ejemplo, en forma de texto libre o una imagen) o pueden ser estructurados en categorías que están representadas consecutivamente o periódicamente mediante el uso de una plantilla, hoja de cálculo o base de datos. La estructura simple de una hoja de cálculo permite a los registros que se representen como filas. Los datos se producen dentro de las celdas formadas por la intersección de filas y columnas que son definidas por metadatos (este término se define más adelante).
4. **Los datos deben ser estandarizados-** La estandarización indica el cumplimiento de un modo ampliamente aceptado de la normalización. Las normas proporcionan términos que definen los datos y las relaciones entre las categorías de datos. Existen dos tipos básicos de las normas, que son indispensables para la gestión de datos biológicos: los metadatos y las ontologías.

5. **Los datos deben ser normalizados**- La normalización trae información contenida dentro de diferentes estructuras en el mismo formato (o estructura). La normalización puede ser tan simple como el uso consistente de un tipo de unidad. La colocación de los datos dentro de una plantilla es un primer paso común a la normalización lo que es un requisito previo para la agregación de datos. Cuando los datos están estructurados y normalizados se pueden movilizar en formatos simples o se puede transformar en otras estructuras. Una de las maneras más comunes para estandarizar y normalizar la información estructurada es mediante el uso de los metadatos: que son términos que definen los datos de manera que puedan servir diferentes a propósitos, tales como ayudar a la gente a encontrar los datos relevantes, es decir que ayudan al descubrimiento de datos, o permiten que los datos se pueden extraer juntos. Los estándares de metadatos definen cómo los datos deben ser nombrados y estructurados, reduciendo así la heterogeneidad de los términos. Las normas pueden definir a los tipos de metadatos que son apropiados para los diferentes tipos de datos. Los conjuntos de términos de metadatos acordados por una comunidad se denominan como vocabularios controlados, uno de los más utilizados en la actualidad es el de Dublin Core (http://dublincore.org/). Mediante la articulación de la aplicación de los metadatos y cómo deben ser formateados, las normas introducen la consistencia que se necesita para la interoperabilidad y el razonamiento de la máquina. Sin estos metadatos, el alcance de las posibles consultas se reduce mucho.
6. **Los datos deben ser atomizados-** La atomización refiere a la reducción de los datos a unidades semánticas mínimas y está en contraste con los datos complejos, tales como imágenes o cuerpos de texto. En las formas atomizadas, pueden existir datos como valores numéricos de las variables, declaraciones binarias o la asociación con términos de metadatos desde vocabularios acordados.

7. **Los datos deben ser publicados-** Se necesita que los investigadores publiquen sus datos mediante una forma que puedan ser más visibles y accesibles. Los científicos pueden publicar en diferentes niveles, en revistas electrónicas, en sitios *Web* institucionales o bien pueden tratar de colocar datos en repositorios centrales. Pero, desgraciadamente sólo una pequeña fracción de los datos se depositan en estos entornos.
8. **Los datos deben ser archivados-** Los datos, una vez publicados, deben ser preservados. No obstante, los proyectos, las iniciativas y las instituciones tienen pocos incentivos para conservar los datos a largo plazo, ya que el proceso incurre en costos generalmente altos, por lo que en muchas ocasiones los repositorios que surgen dentro de los proyectos tienen una esperanza de vida limitada.
9. **Los datos deben ser de libre acceso-** El acceso abierto *(Open Access)* trata sobre el acceso inmediato, sin requerimientos de registro, suscripción o pago (es decir, sin restricciones) a material digital educativo, académico, científico o de cualquier otro tipo, principalmente artículos de investigación científica de revistas especializadas y arbitradas mediante el sistema de revisión por pares o *peer review* (Wikipedia, 2014). El acceso abierto mejora de la captación, uso, aplicación e impacto de la producción de la investigación. En este contexto existen otras convenciones como *Linked Open Data* (Datos abiertos y vinculados) que describe un método de publicación para que los datos puedan ser interconectados y más útiles. Se basa en tecnologías *Web* estándar, tales como HTTP, RDF y los URI, para compartir información de una manera que puede ser leída automáticamente por ordenadores. Esto permite que sean conectados y consultados de diferentes fuentes (Wikipedia, 2014).
10. **Los datos deben ser confiables-** Una vez que se accede a los datos, los consumidores pueden revelar errores y/u omisiones, que pueden corregirse mediante un proceso de curación de datos (se explica más adelante) cada proceso de curación debe ser documentado para ayudar al consumidor a evaluar si la fuente es adecuada para su propósito.

11. **Los datos deben ser atribuidos-** Se requiere cuidado especial al atribuir los datos resultantes de la combinación de uno o más conjuntos existentes de modo que toda la inversión intelectual se acredita adecuadamente.
12. **Los datos deben ser manipulados-** Un valor de contar con grandes cantidades de datos de manera apropiada con anotaciones disponibles en la *Web* es que los usuarios pueden visualizarlos, analizarlos, descargarlos (en algunas ocasiones), ponerles anotaciones o comentarios, etc.
13. **Los datos deben ser registrados y reconocibles-** Se deben mejorar y desarrollar nuevos motores de búsqueda en la *Web,* para revelar el contenido de bases de datos.

Desafortunadamente todavía es raro que los datos creados en la mayoría de las sub-disciplinas se preparen con todos estos requisitos, por lo que es indispensable contar con especialistas que lleven a cabo todas las tareas mencionadas, para conseguir que la información cumpla con los objetivos mencionados con anterioridad, en consecuencia, se ha presentado un cambio importante en la forma en la que se guarda, recupera y maneja la bibliografía.

Las recientes iniciativas para aumentar el acceso, y el apoyo institucional, han comenzado a ayudar a cambiar proporcionando lugares de publicación, donde todo el mundo puede acceder a las publicaciones y sus datos adjuntos libremente y en forma oportuna. Los usuarios de los artículos de libre acceso se benefician mediante la apertura de nuevas vías proporcionadas en bases de datos biológicas, que permiten la continua expansión, un acceso rápido a la literatura publicada y el postularse para la audiencia más amplia posible (Hirschman *et al.*, 2010).

Bases de datos

Con el desarrollo vertiginoso que ha tenido la informática en las últimas décadas, se han desarrollado numerosos procedimientos para sistematizar la información, un ejemplo de esto son las bases de datos y como ya se mencionó en este proyecto sólo se abordarán las bases de datos bibliográficas. Las bases de datos bibliográficas son sistemas informáticos de registros de publicaciones científicas, que sistematizan la información contenida en bibliotecas y/o revistas (Michán y Morrone, 2002) con el propósito de almacenar, mantener y generar información, y que debido al crecimiento exponencial de la

cantidad de datos biológicos que se producen diariamente y a que se requieren medidas revolucionarias para su gestión, análisis y accesibilidad. Las bases de datos en línea se han convertido en importantes vías para la publicación de esos datos (Howe *et al.*, 2008), de esta manera los recursos electrónicos con literatura sobre ciencias que pueden ser consultados vía Internet permiten el acceso inmediato a las colecciones de datos digitales con información generada por los investigadores, que van desde complejos sistemas de información que permiten la consulta simultánea de grandes colecciones bibliográficas (bases de datos de literatura) hasta pequeños directorios de literatura. En la última década, las bases de datos electrónicas han revolucionado la forma en que los científicos acceden a la información. Hoy en día, los repositorios digitales por ejemplo, acomodan grandes cantidades de datos, sirven como almacenamiento de los registros primarios y auxiliares, permiten anotaciones funcionales, y la conservación de la información. Igualmente importante, es que permiten cambios a la información almacenada, para que pueda ser corregida, actualizada, normalizada, estandarizada y revisada (Rzhetsky *et al*, 2008).

Otra de las importantes funciones de las bases de datos, es que éstas representan un instrumento idóneo para documentar satisfactoriamente un acervo de información, lo que es indispensable, ya que, una de las mejores formas de preservar el conocimiento sobre una disciplina es por medio de la documentación. Este proceso consiste en capturar y sistematizar la mayor cantidad de información posible sobre un tema (Michán y Llorente, 2003).

Además de los atributos ya mencionados, las bases de datos nos proporcionan otras ventajas, como las que sugieren Michán y Morrone (2002) en su artículo *"Historia de la taxonomía de coleóptera en México durante el siglo XX: una primera aproximación"* y que se enlistan a continuación:

1. Almacenar mucha información en poco espacio.
2. Sistematizar los datos de acuerdo a las necesidades del proyecto.
3. Tener fácil acceso a la información.
4. Hacer búsquedas con base en distintas entradas.
5. Procesar los datos cuantitativa y cualitativamente.
6. Interrelacionar los resultados utilizando distintas variables.
7. Actualizar rápidamente la información.

8. Hacer conversiones, pues generalmente hay compatibilidad entre distintas bases de datos.

Con el fin de transferir datos de la literatura publicada en una base de datos, se han tenido que (1) desarrollar estrategias y herramientas que permiten conservar estos datos; (2) crear y modificar esquemas de bases de datos para dar cabida a nuevos tipos de datos; y (3) el desarrollo de las páginas *Web* mediante las cuales se accede para buscar y ver los nuevos datos (Hirschman *et al.*, 2010), se requiere también de técnicas de computación avanzada y el aumento de la financiación de organizaciones privadas y gubernamentales.

Para afrontar el reto de captar adecuadamente, almacenar y analizar la gran cantidad de datos presentes en la literatura científica, el número y el alcance de las bases de datos científicas se ha disparado en los últimos años, la creación de un nuevo método de análisis de la información, la biocuración. De hecho, el énfasis reciente en la expansión de recursos computacionales, capaces de gestionar y analizar datos biológicos complejos, presenta una demanda cada vez mayor de especialistas (biocuradores) capaces de interpretar la literatura científica cada vez más compleja y la extracción de datos relevantes de manera eficiente, y consistente, aunque los deberes de estos especialistas en biocuración son diversos (Salimi y Vita, 2006).

A menudo damos por sentado que la riqueza de datos biológicos ya está disponible para consultarlos, como biólogos, a menudo podemos buscar en Internet y encontrar un conjunto de datos que cumpla con nuestras necesidades; sin embargo, estos suelen ser puestos a disposición de los usuarios a través de uno de las muchas bases de datos biológicas que se han creado en los últimos 20 años. Las personas que trabajan con esta información y la ponen en los formatos que nos permiten trabajar fácilmente con ella se especializan en esta disciplina emergente (Bateman, 2010).

Biocuración

La Biocuración puede resumirse como la transformación de los datos biológicos a una forma organizada. Esto se logra sólo a través de los esfuerzos combinados de la comunidad experimental que genera los datos, los biocuradores que organizan los datos y el software y los desarrolladores que hacen que los datos estén disponibles en la *Web* para uso de todos (Bateman, 2010). Sin embargo de manera más

particular la Sociedad Internacional de Biocuración (ISB, por sus siglas en inglés) (http://www.biocurator.org/, 2009-2014) sugiere que la biocuración implica la traducción y la integración de la información biológica relevante dentro una base de datos o en un recurso que permite la integración de la literatura científica y que los objetivos principales de la biocuración son la representación precisa y completa de los conocimientos biológicos, facilitar el acceso a estos datos y "formar" una base para el análisis computacional.

St Pierre y McQuilton, 2009 mencionan que la biocuración, es identificar y recopilar datos, analizar y extraer las conclusiones, incorporar, representar y mostrar los datos en formatos útiles para diferentes tipos de audiencias; esto significa identificar la literatura relevante, traducir los resultados experimentales en un lenguaje estandarizado de búsqueda, trabajar con bases de datos y desarrolladores de sitios *Web* para garantizar la facilidad de uso.

Pero, en atención de este proyecto y sus objetivos, se tomará a la biocuración como: la traducción e integración de la información relevante en biología en una base de datos o en un recurso que permita la integración de la literatura científica, cuyos objetivos se centran en formar conjuntos de datos de gran tamaño, representar de manera precisa los conocimientos biológicos, facilitar el acceso a los datos y generar una base de datos para su análisis computacional (Biocurator.org, 2014)

La biocuración es una tarea que está en el corazón mismo de la biología moderna, muchos de los proyectos biológicos a gran escala hoy, que van desde la creación de repositorios, colecciones, experimentos anotados, vocabularios controlados y ontologías, hasta aquellos que proporcionan evidencias biológicas de la literatura en las bases de datos, utilizan esta práctica. La curación digital puede llegar a ser tediosa, difícil y cara, por lo general, requiere de la participación de varias personas con diversas aptitudes, la productividad, la resistencia, la experiencia, la tendencia a errar, y los prejuicios personales, son algunas de las problemáticas que se pueden presentar. Pero a pesar de sus dificultades y la imprecisión en los resultados, la curaduría es crítica (Rzhetsky *et al.*, 2009).

La biocuración comienza recogiendo o recopilando la literatura de colecciones bibliográficas que pueden o no estar disponibles en la *Web* con el fin de extraer la mayor cantidad de información publicada, para agregarla a las secciones correspondientes de la base, y para conectar las piezas individuales de la información con los datos existentes y así crear una imagen más grande y más

compleja del estudio, investigación o área. Los biocuradores necesitan tener una sólida formación en biología y deben poseer dedicación a los detalles científicos, la curación de literatura efectiva requiere que los curadores sepan qué tipo de información es la que los investigadores necesitan consultar y cómo acceder a esos datos. De hecho, las necesidades de información han cambiado con el tiempo, y los esfuerzos para responder a estas necesidades han ampliado el alcance de la biocuración y las responsabilidades del curador (Hirschman *et al*, 2010).

La naturaleza de los datos pertinentes ha obligado a diversas instancias a establecer pautas de curación bien definidas para promover la coherencia y delinear claramente el objetivo, la representación de los datos y reducir su subjetiva interpretación. A pesar que algunos biocuradores han desarrollado un *Manual de Curaduría;* que proporciona instrucciones detalladas sobre las estrategias y procedimientos para capturar, anotar, y la introducción de datos complejos y detallados de la literatura, su uso no puede ser extendido en todos los casos de biocuración*,* ya que hay situaciones difíciles que surgen inherentemente durante cada uno de ellos, Salimi y Vita (2006) señalan, por ejemplo, que a menudo nos encontramos terminologías inconsistentes en la literatura que presentan retos formidables para nuestra interpretación uniforme de los datos. Los científicos a menudo utilizan nomenclaturas diversas y controvertidas, los métodos utilizados para llevar a cabo un experimento pueden ser un tanto oscuros o contradictorios, las conclusiones de los autores pueden ser difíciles de representar con base en las limitaciones de los campos en las bases de datos y nuestras pautas de curación, tipos de ensayo de nueva creación pueden requerir la interpretación y la asignación a un grupo de ensayo particular. Por lo tanto, cada día aparecen nuevos desafíos para tratar los datos.

Hay una gran cantidad de retos para los investigadores, profesores y alumnos que pretenden realizar una biocuración, ya que, debido al desarrollo en los métodos de investigación, los biocuradores deben adaptarse a la cantidad y a la variedad de datos publicados que crecen exponencialmente cada año, y como mencionan St. Pierre y McQuilton (2009), es esta variedad la que presenta el mayor desafío, ya que es relativamente fácil curar datos que se conocen bien, es mucho más difícil entender y crear bases de datos para los nuevos tipos de datos.

Para garantizar aún más la precisión de la biocuración, existen diversas herramientas informáticas disponibles en la red, por ejemplo: es posible obtener las referencias personales de los autores de los trabajos para ponerse en contacto y aclarar algunos detalles o para solicitar información específica (Salimi y Vita, 2006), se ha iniciado el uso de técnicas analíticas como: la minería de texto (Text Mining) y la semántica para agilizar los procesos de curación de la literatura (Michán, 2011), estás herramientas nos ayudan a resolver algunos problemas técnicos, como la obtención de texto completo en un formato adecuado para la exploración, incluyendo leyendas de las figuras y los datos suplementarios, pueden ser capaces de reconocer con precisión la información más importante y asociarla con diferentes vocabularios controlados (Hirschman *et al*, 2010).

El papel de la biocuración es dinámico y evoluciona en paralelo con la evolución de la bioinformática, y su labor sugiere una reducción de la brecha entre el conocimiento y su accesibilidad; independientemente de la fuente y el sujeto de los datos, el biocurador los agrega a una base de datos y ésta a su vez, a una colección bibliográfica donde se hace más accesible a los investigadores. Los datos organizados y analizados están por tanto, disponibles para la recuperación, el análisis, y la descarga por parte del usuario final en un formato mejorado, que proporciona una dimensión añadida de la utilidad que de otro modo no estaría presente (Salimi y Vita, 2006).

Las personas encargadas de la biocuración (definidas como biocuradores por Salimi y Vita, 2006), tienen que ser capaces de entender los datos científicos e incorporar las directrices de curación de una manera que se mantenga la integridad de estos datos; por ello, la comunicación abierta entre estos y la comunidad científica puede fomentar una mejor comprensión de las dificultades encontradas en la biocuración y facilitar un enfoque más estandarizado para la representación de datos en la literatura. Ante el uso creciente y la contribución a diversas bases de datos, Salimi y Vita (2006) ponen en evidencia la necesidad de establecer o perfeccionar vocabularios y definiciones biológicas estandarizadas. Las nuevas metodologías permiten que mientras los datos científicos crecen a un ritmo exponencial, también lo hace la demanda constante de bases de datos confiables, consistentes y precisas.

Se espera que los modelos de curación alternativos, probablemente relacionados con la participación más directa de la comunidad de investigación, ayuden a abordar la brecha de datos entre la cantidad cada vez mayor de información publicada en la literatura y la cantidad de datos disponibles a través de bases de datos biológicas curadas (Hirschman *et al*, 2010).

Actualmente existen sitios en Internet que ofrecen distintos servicios e información relacionada con la biocuración, por ejemplo: *Digital Curation Center* o DCC (http://www.dcc.ac.uk/) *y The International Society for Biocuration* o ISB (http://www.biocurator.org/home.shtml), esta última además, presenta una liga a una base de datos especializada, DATABASE: *The journal of Biological Datebases and Curation* (http://database.oxfordjournals.org/) donde podemos encontrar documentos referentes a la biocuración.

ANTECEDENTES

El trabajo realizado por la Dra. Layla Michán Aguirre y el Dr. Jorge Enrique Llorente-Bousquets en 2003. *"La taxonomía en México durante el siglo XX"* Publicaciones especiales del Museo de Zoología, Número 12. Facultad de Ciencias, UNAM. 237 pp., en donde se presenta una visión del desarrollo de la taxonomía en México durante el siglo XX; y que para cumplir con sus objetivos diseñaron **"TaXMeXX"**, una base de datos relacional, bibliográfica y taxonómica, representa el antecedente directo de este proyecto, debido a que la parte nuclear del mismo es la realización de un trabajo de curación a esta base de datos.

TaXMeXX entonces, es una base de datos que se diseñó originalmente en el programa Microsoft Office Access 2003®, como una herramienta para la compilación, organización, sistematización, clasificación, representación tabular y gráfica de la información obtenida a partir del análisis de las revistas y los artículos que tratan o contienen información sobre la taxonomía de México durante el siglo XX (del 1° de enero de 1901 hasta el 31 de diciembre de 2000), esta tiene almacenada la información sobre las revistas, los volúmenes y los artículos, los autores, las instituciones, los taxones, la región y el tipo de trabajo taxonómico. La base fue procesada y analizada para incluirse en una narrativa histórica, donde se pretende examinar el desarrollo de la taxonomía en México durante el siglo XX (Michán y Llorente, 2003).

La metodología en la cual se construyó la base de datos fue la siguiente (figura 3): 1) primero se hizo una búsqueda, lo más exhaustiva posible, de las publicaciones que pudieran contener conocimiento taxonómico, 2) de esta búsqueda se encontraron 139 revistas, 3) para posteriormente elegir solo 28 de acuerdo a diferentes criterios, 4) se capturan los metadatos de estas revistas, 5) se continúa entonces con la selección de artículos, 6) para ello se determinó si se trataba de un artículo taxonómico o no, 7) si era considerado como taxonómico se capturaba y se registraban sus metadatos, 8- si se consideraba no taxonómico entonces se eliminaba. Para conocer la metodología completa y los detalles de los criterios de selección, se puede consultar la obra original.

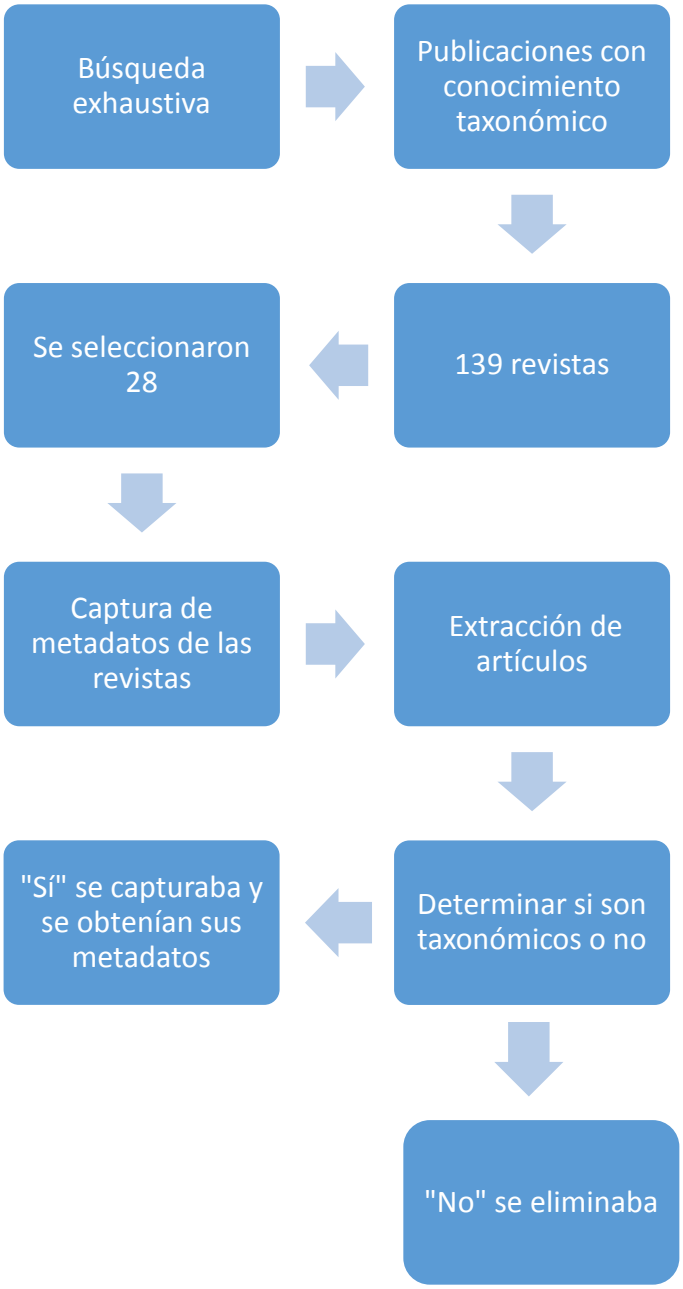


Figura 3: Metodología para la construcción de TaXMeXX

La base de datos TaXMeXX constituye una herramienta que contiene información relevante y permite procesar una muestra significativa de lo publicado sobre la taxonomía en México durante el siglo XX pues incluyó hasta ese momento, el 90% de la información producida en y para esa disciplina. Aunque la base de datos fue construida con una intención histórica también puede aplicarse para consultas bibliográficas, ya que contiene la mayoría de los artículos taxonómicos escritos y publicados en

México durante el siglo pasado, y por lo tanto, permite la extracción de información acerca de los artículos (muchos de ellos ya con resumen y *abstract* indexados y/o con ligas que nos redirigen al texto completo), los autores, las revistas y las instituciones. Esta base de datos permite también la consulta por entidad federativa, es decir, podemos obtener información de los trabajos, autores, revistas e instituciones que publicaron algo relacionado con la taxonomía de alguna o algunas entidades federativas de la República Mexicana por ejemplo, Aguascalientes, Sonora, etc., además también permite la consulta por grupos taxonómicos, ejemplo, reptiles, peces, etc.

Existen otras colecciones en México que contienen y donde podemos consultar la información biológica (incluyendo la taxonómica) de nuestro país, por ejemplo:

- ***Sistema de Información Ciencias, Facultad de Ciencias. UNAM.*** (http://repositorio.fciencias.unam.mx:8080/xmlui). Es un Sistema de Información que además de alojar la base de datos TaXMeXX, contiene otras colecciones importantes:
    - Repositorio de la Facultad de Ciencias.
    - Colección del Dr. Harry Brailovsky (Ramírez-Martínez, D. 2014)
    - Colección de la Revista de la Sociedad Mexicana de Historia Natural (Ramírez-Martínez, D. 2014)
    - Colección de Nuevas Especies Mexicanas (Ramírez-Álvarez, D. 2014)
    - Entre otras.
- ***Red Mundial de Información sobre Biodiversidad*** (REMIB) (http://www.conabio.gob.mx/remib/doctos/remib_esp.html), creada en noviembre de 1993, es un sistema computarizado de información biológica, incluye bases de datos de tipo curatorial, taxonómico, ecológico, cartográfico, bibliográfico, etnobiológico, de uso y catálogos sobre recursos naturales y otros temas. Está red está conformada por centros de investigación y de enseñanza superior, públicos y privados, que poseen tanto colecciones biológicas científicas como bancos de información (CONABIO, 2014).

Por otro lado, han surgido organizaciones internacionales que han generado bases de datos sobre la información biológica a nivel mundial, por ejemplo:

- ***Global Biodiversity Information Facility*** (GBIF) (http://www.gbif.org/) es una organización intergubernamental que nace en 2001 y que comprende en la actualidad 53 países y 43 organizaciones internacionales. Su objetivo es dar acceso, por medio de Internet, de manera libre y gratuita**,** a los datos de biodiversidad de todo el mundo, para apoyar la investigación científica (GBIF, 2014).
- ***Encyclopedia of Life*** (EOL) (http://eol.org/) comenzó en 2007, es ahora una gran base de datos y un portal en línea de acceso libre que reúne información biológica confiable, procedente de fuentes de todo el mundo como museos, sociedades científicas, científicos expertos, etc., (EOL, 2014).

Muchos investigadores enfatizan la importancia de desarrollar, mantener e integrar las colecciones bibliográficas (bases de datos) sobre biodiversidad, que contengan información bibliográfica sobre los taxones de la región, ya que, a diferencia de otras áreas biológicas la literatura taxonómica, no caduca, y cualquier dato publicado sobre alguna especie es de un alto valor biológico (Michán y Morrone, 2002).

Por todo lo expuesto anteriormente, este trabajo forma parte de una línea de investigación del Laboratorio CIIB, FC UNAM en el que se planean, diseñan, publican, mantienen y analizan bases de datos bibliográficas en línea de acceso abierto sobre biodiversidad mexicana. Una de estas bases de datos es TaXMeXX, que fue curada durante el desarrollo de este proyecto.

OBJETIVOS

General:

Realizar la biocuración de los artículos analizados (6,152 registros) que contiene la base de datos TaXMeXX.

Particulares:

1. Normalizar los metadatos analizados correspondientes a la colección bibliográfica TaXMeXX.
2. Actualizarla información por medio de la recuperación de los artículos analizados en TaXMeXX.
3. Atomizar el contenido de los artículos analizados en TaXMeXX.
4. Agregar anotaciones de títulos, autores, *abstract*, resumen, sumario, *summary*, sinopsis e idioma.
5. Completar la información pertinente a abreviaturas de autores y lenguaje para estándares de Dublin Core.
6. Actualizar los URL´s de artículos disponibles en otras bases de datos en línea.
7. Publicar la base de datos curada en el Sistema de Información Ciencias, UNAM.

EXPOSICIÓN DEL TEMA

La biocuración, es un proceso que consta de una serie de etapas. De acuerdo al Centro de Curación Digital (DCC, http://www.dcc.ac.uk/, 2014), las etapas pueden estar representadas en un ciclo donde se sobreponen o traslapan unas con otras. Este proyecto se enfoca en una de ellas llamada reevaluación, específicamente, en un proceso interno que podríamos definir como evaluación, selección y desecho de información, pero para fines de este proyecto lo denominaremos normalización ó curación. Por lo tanto la normalización es el proceso de organizar de manera eficiente los metadatos, y sus objetivos son: la eliminación de los datos redundantes, evitar problemas de actualización de los datos en las tablas, garantizar que las dependencias entre ellos sean lógicas y transformar datos complejos en conjuntos de estructuras más pequeñas (Thessen y Patterson, 2011).

Para alcanzar los objetivos de este trabajo, se aplicaron las siguientes herramientas informáticas:

- Consulta en línea de sistemas de información, colecciones, índices, catálogos y repositorios bibliográficos (Tabla 1)
- Manejador de bases de datos en formato .accdb (Microsoft Office Access 2010®)
- Repositorio (DSpace)
- Estándar de metadatos Dublin Core®
- Hoja de cálculo en formato .xls (Microsoft Office Excel 2010®)
- Acrobat Reader® versión 10.9.22
- Reconocimiento óptico de caracteres (OCR)
- Mendeley como manejador de bibliografía en línea
- Google Drive para realizar distintas tareas y anotaciones

Tabla 1: Fuentes electrónicas consultadas

Nombre	Descripción	Dirección electrónica
Periódica	Índice de revistas latinoamericanas en ciencias (Base de datos bibliográfica)	http://132.248.9.1:8991/F/S588FAA176Y2T969MK5Q47E2YJDD7MNBHA7U66CYI1L28VHJJC-20791?func=find-b-0&local_base=per01
Latindex	Sistema regional de información en línea para revistas científicas de América Latina, el Caribe, España y Portugal	http://www.latindex.unam.mx/#
Ulrich's	Fuente de información bibliográfica de publicaciones periódicas	http://ulrichsweb.serialssolutions.com/
Web of Science	Base de datos bibliográfica	http://apps.webofknowledge.com/UA_GeneralSearch_input.do?product=UA&SID=4FfsmZ57YKC2lrsajt9&search_mode=GeneralSearch
Biblat	Portal especializado en revistas científicas y académicas publicadas en América Latina y el Caribe	http://biblat.unam.mx/es/
Redalyc	Red de revistas científicas de América Latina y el Caribe, España y Portugal (sistema de información científica)	http://www.redalyc.org/home.oa
Scielo	Biblioteca científica electrónica en línea	http://www.scielo.org/php/index.php?lang=es
Google académico (Google scholar)	Buscador de Google especializado en artículos de revistas científicas	http://scholar.google.es/
E-journal (sólo actualizado hasta el 2010). El portal actual es: Revistas UNAM	Colección, en formato digital, de revistas científicas y arbitradas de la UNAM	http://www.ejournal.unam.mx/ http://www.revistas.unam.mx/
Sistema de Información, F. Ciencias, UNAM	Repositorio institucional de literatura académica	http://repositorio.fciencias.unam.mx:8080/xmlui
Biblioteca CCG-IBT UNAM	Biblioteca de revistas electrónicas	http://biblioteca.ibt.unam.mx/revistas.php
Sitio web de cada revista (si es que lo hubiera)		

El protocolo de trabajo consistió de 7 etapas y se esquematiza en la Figura 4.

1. Descargar los 6,152 registros correspondientes a los artículos analizados contenidos en la base de datos TaXMeXX, que está depositada en el Sistema de Información Ciencias, UNAM, en el programa Access (Microsoft Office 2010®)
2. Seleccionar los primeros 18 campos de los 39 presentes en la base de datos (Tabla 2).
3. Agregar nuevos campos a la parte seleccionada
4. Definir cada campo o metadato.
5. Realizar una búsqueda lo más exhaustiva posible de cada uno de los artículos en fuentes disponibles en la red, comparar la información contenida en la base de datos con respecto a la localizada y realizar las acciones correspondientes.
6. Indexar toda la información a la base de datos.
7. Crear una lista con las URL'S de los artículos encontrados.

Por lo tanto, para esta investigación y en atención a las necesidades del proyecto, primeramente, se eligió Microsoft Access 2010® como manejador de la base de datos, ya que es una herramienta que permite almacenar, organizar, buscar y presentar información de una manera fácil y dinámica, cuenta con asistentes y herramientas de diseño para crear formularios para la captura de información, así como reportes e informes, incluyendo diversos tipos de gráficos, lo cual lo hace un manejador sumamente útil, versátil y sencillo de usar. La base de datos TaXMeXX cuenta con un total de 8,275 artículos, pero que bajo los criterios de los autores, sólo se analizaron 6,152 dando como resultado el mismo número de registros y son los que, por su importancia y la información contenida, se curaron.

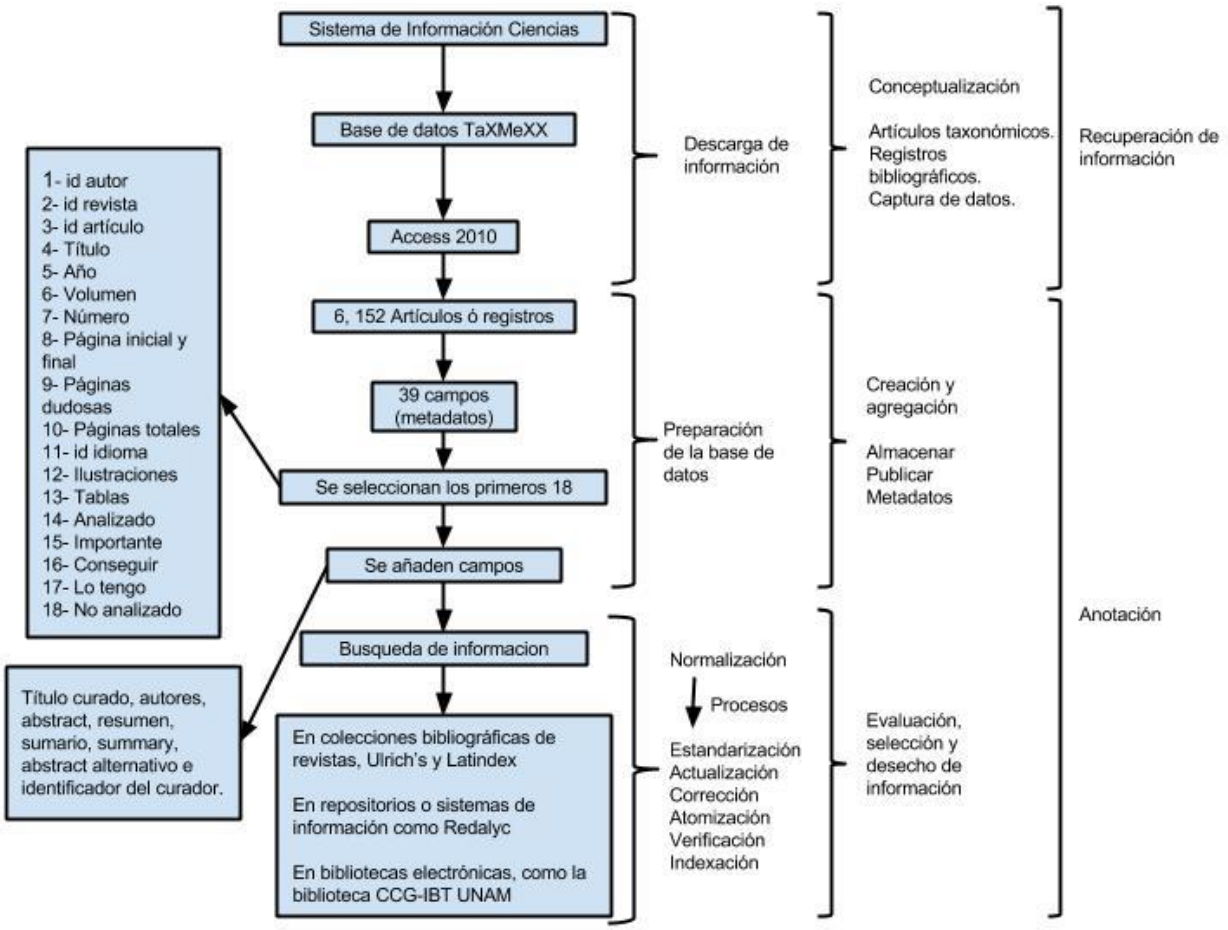


Figura 4: Proceso de curación de la base de datos TaXMeXX

La base ya descargada en Microsoft Access 2010® guarda una relación tanto vertical (columnas) como horizontal (filas), la relación vertical está definida por los campos o metadatos. Un campo es un espacio de almacenamiento donde se escribe la información relacionada con un metadato en particular (RAE, 2014) y cada campo tiene asignada una etiqueta (esquema de metadatos Dublin Core®) que lo distingue y sirve para homogeneizar la información. La relación horizontal está dada por las filas, cada una de ellas es un registro, éste a su vez se compone de varios campos o metadatos que representan toda la información obtenida de un objeto, en este caso artículos taxonómicos. Cada registro en la base de datos TaXMeXX está representado por los 39 campos o metadatos mostrados en la Tabla 2.

Tabla 2. Campos o metadatos presentes en la base de datos TaXMeXX

1.Id autor	2.Id titulo	3.Id revista
4.Título	5.Año	6.Volumen
7.Número	8.Página de inicio y final	9.Páginas dudosas
10.Total de paginas	11.Id idioma	12.Ilustraciones
13.Tablas	14.Analizado	15.Importante
16.Conseguir	17.Lo tengo	18.No analizado
19.Histórico no taxonómico	20.No taxonómico	21.Citas CONABIO
22.Notas 2	23.Número de páginas juntos	24.Citas
25.Referencia completa	26.Referencia	27.Datos cita completos
28.Filtro comodín	29.Especies descritas	30.Resumen
31.Expr1015	32.Palabras clave	33.URL
34.Expr1019	35.UT	36.UTID
37.TIO	38.Autores juntos	39.Expr1027

La base de datos fue dividida seleccionando los primeros 18 metadatos. Posteriormente se le agregaron los siguientes campos:

1. Título curado: Contiene el título del artículo ya revisado y si fuese el caso con las correcciones o modificaciones hechas.

2. Autores: Contiene los nombres de todos los autores del artículo en un mismo campo, consecutivos y separados por una doble barra, para normalizar la información, se escribirá el nombre completo de cada autor y todas sus derivaciones, esto es para evitar duplicaciones y confusiones.

3. Resumen: Aquí se colocó el resumen en español del artículo si es que lo presenta. Debido a la variedad de documentos analizados, algunos de ellos presentan sumario o sinopsis e lugar de un resumen, por lo que éstos también fueron capturados en este campo.

4. *Abstract*: Aquí se colocó el *abstract y/o summary* del artículo si es que lo presenta, cabe mencionar que en la base de datos se capturaron documentos en 6 idiomas {Español (es), Inglés

(en), Francés (fr), Italiano (it), Portugués (pt) y Alemán (de)} por lo que algunos de ellos presentan resumen en otro idioma diferente al Español e Inglés, en este campo se capturó esa información.

5. Identificador del curador: se trata de un número consecutivo del 1 hasta el 6,152 para llevar el control de lo realizado en este proyecto.

Seguido esto, se define cada metadato, para saber qué información contiene o está presente en cada campo y de esta manera reconocer la información que puede ser indexada o modificada. Finalmente, se realizó una búsqueda exhaustiva de cada uno de los artículos analizados, en diversas fuentes electrónicas disponibles en la *Web* (mencionadas en la Tabla 1) de esta manera se obtuvo la información necesaria para compararla con la contenida en TaXMeXX, y así, reconocer los errores y corregirlos. Por último, se hizo uso de un servicio OCR gratuito en línea (http://www.newocr.com/), para extraer el texto de todos aquellos artículos en formato PDF que no permitieran copiar directamente los resúmenes, *abstracts*, sumarios, *summary* y sinopsis, para agregarlos a la base de datos y de este modo ahorrar muchas horas de trabajo y esfuerzo de transcripción. Simultáneamente se usó una hoja de cálculo Microsoft Excel 2010® para crear una lista de las URL'S (direcciones electrónicas) de los artículos encontrados en su formato completo, relacionándolas con los identificadores: id artículo e id revista, además del año, para su posterior identificación, localización y utilización.

Es importante mencionar que los identificadores; id autor, id revista e id artículo, consisten en un número único dentro de la base que lo relaciona directamente con el catálogo de nombres de autores, revistas y artículos respectivamente.

Anotaciones

Al concluir la curación de la totalidad ó 100% de la base de datos TaXMeXX, que consta de 6,152 artículos o registros analizados, se encontró que a partir de este trabajo de normalización se ha indexado información en algunos campos (Tabla 3).También se localizaron errores específicos y se corrigieron realizando un total de 3,010 cambios pertinentes, que se muestran en la Tabla 4 y se describen en las gráficas 1 y 2.

Si tomamos en cuenta que la base de datos tiene tanto relaciones horizontales (6,152 registros) como verticales (campos o metadatos), pero debido a que no se hicieron modificaciones en algunos de estos últimos (Id artículo, id revista, id autor, analizado, importante, conseguir, lo tengo y no analizado) por tratarse de identificadores propios de la base, así como campos que no pueden ser curados bajo los criterios de este trabajo, quedaron un total de 11 campos en los que se centró la normalización (año, volumen, número, página de inicio y final, páginas dudosas, páginas totales, id idioma, ilustraciones, tablas, título y autor), eso dio como resultado un total de 67,672 datos bibliográficos curados, de los cuales en 7,410 se hizo alguna modificación. Esto representa el 10.95 % del total de la base.

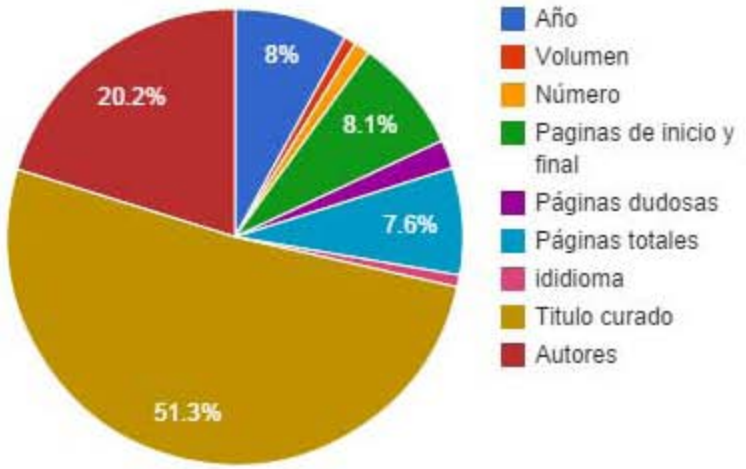
Tabla 3. Información indexada a la base de datos

Campo o metadato	**Antes de la curación**	**Después de la curación**
Ilustraciones	84	2,208
Tablas	46	2,192
Resumen	0	1,111
Abstract	0	914
Summary	0	315
Abstract alternativo (idioma diferente al inglés)	0	11
Sinopsis	0	2
URL (direcciones electrónicas donde se encuentran los artículos en formato completo)	0	2,040 (que representa el 33.16 % del total de la base)
Total	**130**	**8,793**

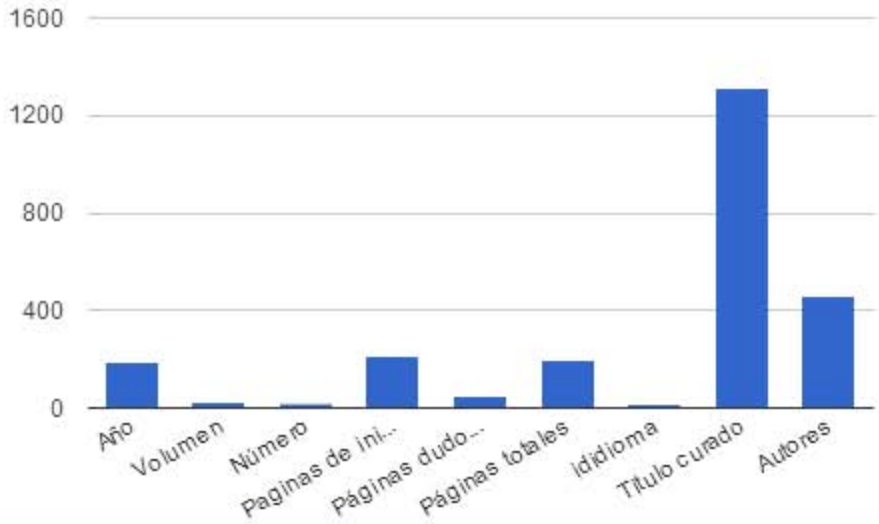
Tabla 4. Cantidad de cambios realizados a los campos en la base de datos TaXMeXX

Campo o metadato	Total de cambios	Porcentaje con respecto al total	Total acumulado
Año	237	7.87 %	237
Volumen	25	0.83 %	262
Número	33	1.09 %	295
Páginas de inicio y final	240	7.97%	535
Páginas dudosas	60	1.99 %	595
Páginas totales	226	7.51%	821
Id idioma	25	0.83 %	846
Título curado	1535	51.00 %	2381
Autores	629	20.90%	3010
Total	3010	100 %	

Gráfica 1. Porcentajes de cambios por dato bibliográfico



Grafica 2. Número de cambios por dato bibliográfico



El trabajo de normalización se aplicó en varios campos de la base de datos TaXMeXX, como se describe a continuación:

Campo Título: este se comparó con la información encontrada en la *Web* (artículo o referencia) para revisarlo y si fuera necesario corregirlo con base en los siguientes criterios:

1. Corregir cualquier modificación, completar o eliminar palabras, frases, etc., que no aparecieran en el título original de la obra.
2. Corregir los errores ortográficos con excepción de aquellos que aparecieran en el título original del documento.
3. Las letras en mayúsculas sólo serán usadas al inicio del título y en los nombres propios.
4. Reducir lo más posible las abreviaciones.
5. No usar punto final.
6. Ningún cambio debe alterar el contexto y significado del título.
7. Los títulos revisados y corregidos se agregaron al campo “título curado”, permaneciendo entonces el título anterior con la etiqueta del metadato título.

El campo autor: Se comparó inicialmente la información contenida en la base de datos con respecto a la encontrada en la red y se realizaron las modificaciones pertinentes respetando la información del documento original si es que éste se encontraba. Después se indexó la información en el campo autor. Si se trataba de más de uno, estos fueron separados por una doble barra, ya que de esta manera la plataforma DSpace (en donde se encuentra el repositorio y donde será publicada TaXMeXX) reconoce a los distintos autores de un registro. Por último, el autor puede tener diversas variantes de su nombre, por ejemplo:

El nombre Jorge Enrique Llorente Bousquets puede aparecer como autor de las siguientes formas:

- Llorente, J.
- Llorente, J. E.
- Llorente-Bousquets, J.
- Llorente-Bousquets, J. E.

Por esta razón se decidió escribir su nombre completo en un campo y en otro alguna de sus variantes para evitar confusiones, para completar la información en el catálogo de autores se agregaron todas las variantes restantes.

En el caso del idioma: la base de datos TaXMeXX hace uso un identificador (id idioma) para señalar el idioma del artículo registrado, el cual funciona de la siguiente manera:

Tabla 5: Identificador de cada idioma presente en TaXMeXX.

identificador	1	2	3	4	5	6
idioma	Español	Inglés	Francés	Alemán	Portugués	Italiano

Pero debido a que esto puede prestarse a algunas confusiones, además del identificador se añadió un código de idiomas según la **ISO 639-1** la cual consiste en usar 204 códigos de dos letras para identificar los idiomas principales del mundo (Wikipedia, 2014). Por lo que la información en TaXMeXX quedó:

Tabla 6: Código de idiomas que se usó en TaXMeXX.

código	es	en	fr	De	pt	it
idioma	español	inglés	francés	Alemán	portugués	italiano

Para concluir la normalización se cambiaron las etiquetas de los metadatos, siguiendo el modelo establecido por Dublin Core® (http://dublincore.org/) que es una organización dedicada a fomentar la adopción extensa de los estándares interoperables de los metadatos y a promover el desarrollo de los vocabularios especializados. Este modelo propone 15 definiciones semánticas para describir los campos, estas definiciones son opcionales, se pueden repetir y pueden aparecer en cualquier orden.

Con respecto a la atomización y al valor agregado a la base de datos TaXMeXX, se hizo lo siguiente: Se agregaron campos a la base de datos donde se indexó la información del resumen y el *abstract* de los artículos que lo tenían. Sin embargo, en el desarrollo del trabajo se localizaron otro tipo de resúmenes: sumario, sinopsis, *summary* y resúmenes en otro idioma. Esta información se capturó separadamente para cada uno de ellos, pero a pesar de que el esquema de Dublin Core® permite generar más etiquetas para realizar una atomización más fina, se decidió no hacerlo ya que esto le restaría interoperabilidad con otras bases de datos que no lo tuvieran de esa manera. Por ello se incluyeron en sólo dos campos: resumen y *abstract.*

Ilustraciones y tablas: Para realizar la curación de estos dos campos en primer lugar se deben establecer los criterios de normalización, esto es, definir o tratar de conceptualizar lo que se considerado como ilustración y lo que se tomará como tabla.

- Ilustración: cualquier imagen, figura, dibujo, fotografía, lámina, gráfica, etc., que fuera mencionada en el texto, o bien que tuviera el pie de figura o título, excepto para las fotografías en las semblanzas y obituarios que raramente contienen estos datos (por ello siempre se contabilizaron)
- Tabla: cualquier tabla o cuadro que tuviera pie de tabla o título adjunto mencionándolo.

Dificultades presentadas y criterios establecidos:

1. **Una ilustración contenía varias figuras** en forma de incisos o números, ejemplo: figura 3 con incisos a, b, c,... o bien lámina III con figuras 1,2,3,...etc., por ello se tomaron los títulos o pies de página para declarar si se trataba de una o varias ilustraciones.
2. **Las imágenes que no tenían título o pie de figura** sólo se contabilizaron si en alguna parte del artículo se les mencionara, a excepción de las semblanzas y obituarios, dado que algunos de ellos anexaban la fotografía del investigador, y por tratarse de una parte importante del artículo éstas si fueron tomadas en cuenta.
3. **Dentro del artículo se presentaban imágenes, pero éstas carecían de información y tampoco se mencionaban en el texto,** por lo tanto no se tomaron en cuenta.
4. **Las ilustraciones que estaban en hojas anexas al artículo** sólo se contabilizaban si estas ilustraciones se mencionaban en alguna de las parte del texto, excepto para las semblanzas y obituarios.
5. **Se mencionan figuras, pero éstas no aparecen** como están mencionadas en el artículo, éstas se contabilizaron tomando en cuenta la información del texto.
6. **No se consideró la información presente en apéndices del artículo** ya que no tienen la palabra tabla o cuadro en su título.
7. **Los mismos criterios fueron aplicados para el campo tablas.**

La información obtenida e indexada después de la curación de los registros se presenta y es comparada con la información contenida antes de la normalización en las tablas 7 y 8 y gráficas 3 y 4.

Tabla 7: Ilustraciones y tablas contenidas en la base de datos antes de la curación

ANTES DE LA CURACIÓN	Número de datos	% con respecto al total de registros
Ilustraciones	84	1.33
Tablas	46	0.75
Total entre las dos	130	
Total de registros	6152	100

Gráfica 3: Ilustraciones y tablas contenidas en la base de datos antes de la curación

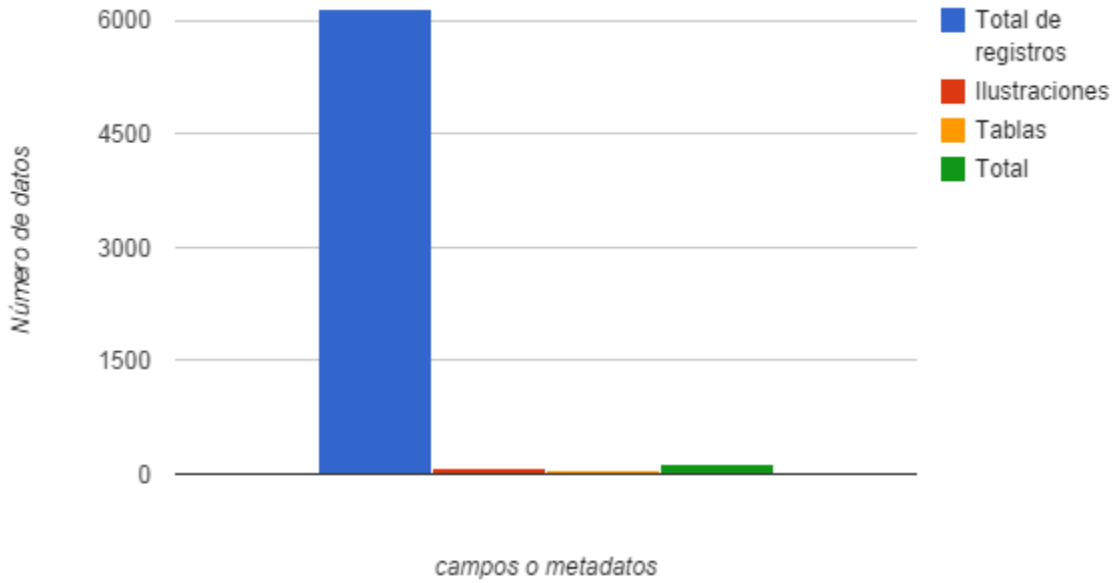
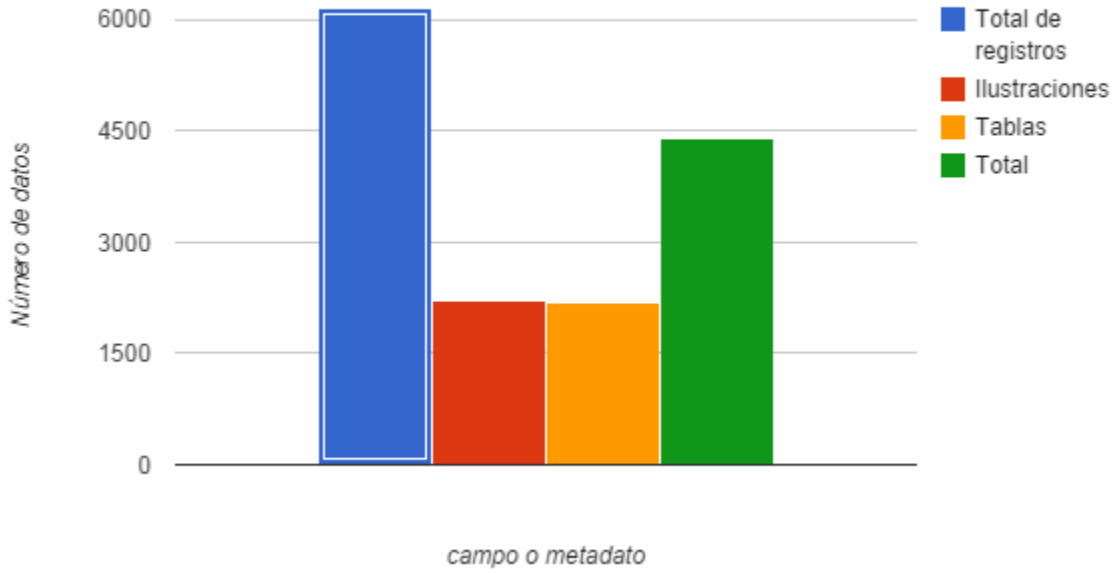


Tabla 7: Ilustraciones y tablas contenidas en la base de datos después de la curación

DESPUÉS DE LA CURACIÓN	Número de datos	% con respecto al total de registros
Ilustraciones	2208	35.9
Tablas	2192	35.63
Total entre los dos	4400	
Total de registros	6152	100

Gráfica 4: Ilustraciones y tablas contenidas en la base de datos después de la curación



Otro resultado importante es la localización de 35 casos de registros repetidos o duplicados. Se tomaron como registros duplicados aquellos que tenían el mismo título y autor, con base en estos criterios se encontraron registros en los que 2, 3 e incluso 4 artículos estaban referidos una misma revista pero en diferentes años ó números, o bien los artículos se encontraban en diferentes revistas (Tabla 5).

Tabla 5: Registros duplicados o repetidos.

Id artículo	Id revista	Año	Titulo	Autor
897	23	1967	Comentarios sobre el concepto moderno de especie y su aplicación a los protozoarios ciliados.	López-Ochoterena, E.
7763	23	1991	Comentarios sobre el concepto moderno de especie y su aplicación a los protozoarios ciliados	López-Ochoterena, E.
4898	102	1985	Contribución al conocimiento de los hongos que crecen en la región de El Texcal, estado de Morelos	Portugal, D. ‖ Montiel, E. ‖ López, L. ‖ Mora, V. M.
4899	102	1985	Contribución al conocimiento de los hongos que crecen en la región de El Texcal, estado de Morelos	Portugal, D. ‖ Montiel, E. ‖ López, L. ‖ Mora, V. M.
4973	13	1979	Discusión sobre Echinofossulocactus	Meyrán, J.
3394	13	1979	Discusión sobre Echinofossulocactus	Meyrán, J.
1285	12	1969	El itinerario y las colectas de Sessé y Mociño en México	McVaugh, R.
7646	21	1974	El itinerario y las colectas de Sessé y Mociño en México	McVaugh, R.
2022	13	1992	El maguey (Agave spp.) y los tepehuanes de Durango	González, M. ‖ Galván, R.
3528	13	1992	El maguey (Agave spp.) y los tepehuanes de Durango	González, M. ‖ Galván, R.
1038	28	1932	Estudio crítico sobre clasificación de las salmonelas	Zozaya, J. ‖ Varela, G.
7733	28	1932	Estudio crítico sobre clasificación de las salmonelas	Zozaya, J. ‖ Varela, G.
7754	23	1991	Félix Dujardín y su "Histoire Naturelle des Zoophytes, Infusoires", 1841	Beltrán, E.
1004	23	1941	Félix Dujardín y su "Histoire Naturelle Des Zoophytes. Infusoires", 1841	Beltrán, E.
899	23	1970	Historia de las investigaciones sobre protozoarios de vida libre de México	López-Ochoterena, E.
7766	23	1991	Historia de las investigaciones sobre protozoarios de vida libre de México	López-Ochoterena, E.
1570	8	1979	La contribución de C. G. Ehrenberg al conocimiento de los protozoarios de vida libre de México	López-Ochoterena, E. ‖ Madrazo, M.
906	23	1991	La contribución de C. G. Ehrenberg al conocimiento de los Protozoarios de vida libre de México	López-Ochoterena, E. ‖ Madrazo, M.
746	23	1943	La deuda de la protozoología con Gary N. Calkins (1869-1943)	Beltrán, E.
7756	23	1991	La deuda de la protozoología con Gary N. Calkins (1869-1943)	Beltrán, E.

Tabla 5: Continuación

898	23	1968	La protozoología dentro de la biología actual	López-Ochoterena, E.
7764	23	1991	La protozoología dentro de la biología actual	López-Ochoterena, E.
1216	12	1952	Las cactáceas del Valle de México	Gold, D. B.
3004	13	1955	Las cactáceas del Valle de México	Gold, D. B.
1229	12	1954	Las Compuestas del Valle Central de México	Paray, L.
1246	12	1958	Las compuestas del Valle Central de México	Paray, L.
1202	12	1953	Las compuestas del Valle Central de México	Paray, L.
1205	12	1949	Las pseudotsugas de México	Martínez, M.
4137	5	1949	Las pseudotsugas de México	Martínez, M.
4393	24	1975	Literatura sobre lepidópteros mexicanos	Beutelspacher, G. L.
4398	24	1975	Literatura sobre lepidópteros mexicanos	Beutelspacher, G. L.
4405	24	1976	Literatura sobre lepidópteros mexicanos	Beutelspacher, M. G.
4411	24	1976	Literatura sobre lepidópteros mexicanos	Beutelspacher, G. L.
1045	23	1947	Lorande Loss Woodruff (1879-1947), miembro honorario de la Sociedad Mexicana de Historia Natural, y sus investigaciones protozoológicas	Beltrán, E.
7757	23	1991	Lorande Loss Woodruff (1879-1947), miembro honorario de la Sociedad Mexicana de Historia natural, y sus investigaciones protozoológicas	Beltrán, E.
751	23	1950	Los géneros de amibas parásitas	Beltrán, E.
7760	23	1991	Los géneros de amibas parásitas	Beltrán, E.
7753	23	1991	Maynard M. Metcalf, su obra científica y el conocimiento de los protociliados	Beltrán, E.
755	23	1940	Maynard M. Metcalf, su obra científica y el conocimiento de los protociliados	Beltrán, E.
2025	13	1991	Nota sobre la distribución de Disocactus ramulosus en México	Lomelí, J. A.
5496	13	1991	Nota sobre la distribución de Disocactus ramulosus en México	Lomelí, J. A.

Tabla 5: Continuación

3772	102	1991	Nuevos registros de poliporáceos estipitados de Jalisco	Vázquez, L. S. ‖ Guzmán-Davalos, L.
4922	102	1991	Nuevos registros de poliporáceos estipitados de Jalisco	Vázquez, L. S. ‖ Guzmán-Davalos, L.
5212	19	1952	Orquídeas mexicanas	Balme, J.
1208	12	1950	Orquídeas mexicanas	Balme, J.
908	23	1977	Panorama retrospectivo de la protozoología mexicana (1841-1986)	López-Ochoterena, E. ‖ Madrazo, M.
7768	23	1991	Panorama retrospectivo de la protozoología mexicana (1841-1986)	López-Ochoterena, E. ‖ Madrazo, M.
4022	5	1966	Plantas nuevas de México	Matuda, E.
4159	5	1965	Plantas nuevas de México	Matuda, E.
3933	5	1943	Registros de géneros y especies nuevos	Caballero, E. ‖ Cerecero, M. C.
5059	5	1942	Registros de géneros y especies nuevos	Caballero, E. ‖ Cerecero, M. C.
5633	4	1982	Restos pleistocénicos de dos especies de Microtus (Rodentia: Muridae), del norte de San Luis Potosí, México	Álvarez Solórzano, T. ‖ Polaco, O. J.
6754	4	1982	Restos pleistocénicos de dos especies de Microtus (Rodentia: Muridae), del norte de San Luis Potosí, México	Álvarez Solórzano, T. ‖ Polaco, O. J.
5381	5	1958	Revisión del género Neodawsonia	Bravo, H. ‖ MacDougall, T.
3073	13	1959	Revisión del género Neodawsonia	Bravo, H. ‖ MacDougall, T.
7824	23	1991	Richard B. Goldschmidt (1878-1958) Zoólogo, geneticista, evolucionista	Beltrán, E.
754	23	1959	Richard B. Goldschmidt. 1878-1958. Zoólogo, geneticista, evolucionista	Beltrán, E.
745	23	1942	Robert Hegner (1880-1942); el hombre, el parasitólogo y el naturalista	Beltrán, E.
7755	23	1991	Robert Hegner (1880-1942): el hombre, el parasitólogo y el naturalista	Beltrán, E.
7895	98	1946	Sobre la no existencia del ciprés Cupressus thurifera H. B. K	Martínez, M.
1195	12	1947	Sobre la no existencia del ciprés Cupressus thurifera H. B. K	Martínez, M.

Tabla 5: Continuación

1467	5	1956	Una nueva especie de Cereus	Bravo, H.
3038	13	1957	Una nueva especie de Cereus	Bravo, H.
5353	5	1957	Una nueva especie de Mammillaria	Bravo, H.
3061	13	1958	Una nueva especie de Mammillaria	Bravo, H.
1559	8	1979	Una nueva especie del género Prepona Boisduval (Lepidóptera: Nymphalidae) de México	Beutelspacher, C.
1621	8	1981	Una nueva especie del género Prepona Boisduval (Lepidóptera: Nymphalidae) de México	Beutelspacher, C.
7758	23	1991	Notas de historia protozoológica. I. - Descubrimiento de los sarcodarios y los trabajos de F. Dujardin.	Beltrán, E.
97	23	1948	Notas de historia protozoológica, I. - Descubrimiento de los sarcodarios y los trabajos de F. Dujardin	Beltrán, E.
1796	8	1978	Nuevos registros de la familia Arctiidae (Lepidóptera) para México	Beutelspacher, C.
2419	8	1987	Nuevos registros de la familia Arctiidae (Lepidóptera) para México	Beutelspacher, C.
7446	13	1973	Thomas MacDougall	Bravo, H.
2362	21	1973	Thomas MacDougall	Bravo, H.

Por último, como observación, la revista Orquídea está bien representada en su base de datos, por lo que se encontraron los 402 dos registros para ésta y se obtuvo la misma cantidad de URL's, sin embargo, estas direcciones electrónicas caducaron y ya no es posible accesar para revisar su contenido, por lo que no fue posible extraer los resúmenes y *abstracs*.

CONCLUSIONES

1. Las bases de datos bibliográficas digitales son una herramienta eficiente en la conservación, presentación y difusión del conocimiento en la biología: De acuerdo al trabajo desarrollado en este proyecto, se planteó que ante la cantidad cada vez mayor de datos generados en la biología, se ha incrementado el uso de la literatura digital, como una estrategia para publicar, difundir y mantener el conocimiento, pero para cumplir con estos objetivos se requiere de la búsqueda y utilización de herramientas electrónicas disponibles en la Web, se propone entonces el uso de bases de datos bibliográficas digitales como herramientas que permiten mantener o conservar la información obtenida de la investigación biológica, para manejarla, presentarla, sistematizarla, actualizarla, etc. Una de estas bases de datos es TaXMeXX. Esta es una base de datos de gran valor taxonómico, que como ya se mencionó anteriormente, cuenta con más del 90% de lo publicado sobre esta área en el siglo XX, y no sólo se trata de artículos relacionados con la taxonomía, sino que también contiene información de los investigadores, las instituciones y las revistas que dedicaron su trabajo a la ampliación del conocimiento taxonómico. Además, esta base de datos bibliográfica presenta una bifuncionalidad tanto taxonómica como histórica. Debido al ritmo acelerado de creación de la información, ésta como otras bases de datos digitales, quedarían obsoletas en poco tiempo, por ello, la importancia de la biocuración como un método con el que se puede actualizar, corregir y homogeneizar la información, además de hacerla accesible para investigadores, profesores, alumnos o cualquier persona interesada en la biología.

2. La biocuración es actualmente una actividad muy importante y se proyecta como una disciplina emergente en la biología. La biocuración, hoy en día, juega un papel crítico dentro de la biología, ya que las nuevas generaciones de biólogos han crecido con el uso de las computadoras, Internet y los dispositivos electrónicos, por lo que, prácticamente no hay algún aspecto de sus actividades en la actualidad, en la que se pueda aspirar, sin la ayuda de las herramientas informáticas, a buscar la información necesaria para la toma de decisiones, de hecho, todos los avances en informática y las disciplinas que convergen en ella, han cambiado la manera de buscar y obtener información, así como nuestra forma de comunicarnos e interactuar,

por lo que se requiere de métodos como la biocuración para el manejo, producción y diseminación de datos e información relevante. En la actualidad, todos los biólogos deben tener la capacidad para buscar, seleccionar, analizar y utilizar estos datos e información. La biocuración hace que estén disponibles con una mejor calidad.

3. Para que el conocimiento pueda cumplir con todos sus propósitos es necesario establecer mejores pautas: La gestión de la información en busca del conocimiento no sólo comprende la tenencia de tecnología de punta (*state of the art technology*), sino que deben conjugarse otros factores en todos los niveles, como son las políticas acertadas de aprendizaje y la creación de grupos interdisciplinarios. Es un hecho que en el proceso de biocuración deben interactuar y participar distintas áreas tanto de la informática como de la biología.

4. La biocuración debe ser realizada por biólogos: El trabajo del biocurador no es sencillo, ya que requiere de una formación como biólogo para poder abordar estos retos. El biocurador se tiene que enfrentar a enormes acervos de registros (cientos o miles) cada día; los datos son más complejos, y aún mucha información no está digitalizada, lo que hace que esta labor sea más complicada. El biocurador también tiene que conocer y usar de manera eficiente las herramientas electrónicas disponibles, le es indispensable el trabajo en equipo, ya que, la biocuración consiste en un trabajo multidisciplinario que puede abordar distintas áreas de la biología.

REFERENCIAS

Bateman, A. (2010). Curators of the world unite: the International Society of Biocuration. *Bioinformatics (Oxford, England)*, 26 (8), p.991. [En línea]. Disponible en: doi:10.1093/bioinformatics/btq101 [Consultada: 25 July 2014].

Biocurator.org. (2014). International Society for Biocuration. *Biocuration*. [En línea]. Disponible en: http://www.biocurator.org/

Castellanos Morales, M. (2013).*Line@: Base de Datos de Colecciones Bibliográficas para Investigación sobre Biodiversidad.*(Tesis de Licenciatura). Facultad de Ciencias. Universidad Nacional Autónoma de México.

Castro, A., Olivares, S. S., Alonso, J. O. y Ramírez, M. E. (2005). Algunas reflexiones sobre la revista electrónica en la UNAM. *Revista Digital Universitaria*, 6 (4).Publicado en línea. http://www.revista.unam.mx/vol.6/num4/art37/abr_art37.pdf

DATABASE. (2014). *The journal of Biological Datebases and Curation.* [En línea]. Disponible en:http://database.oxfordjournals.org/

Dcc.ac.uk. (2011). *What is digital curation?* | *Digital Curation Centre*. [En línea]. Disponible en: http://www.dcc.ac.uk/digital-curation/what-digital-curation

Dublin Core® (2014). Dublin Core® Metadata Initiative. [en línea]. Disponible en: http://dublincore.org/

eBird. (2014). Portal de registros de aves. [En línea]. Disponible en: www: //ebird.org

EOL. (2014).Encyclopedia of Life. [En línea]. Disponible en:http://eol.org/

GBIF. (2014). Global Biodiversity Facility. [En línea]. Disponible en: http://www.gbif.org/

Hirschman, J., Berardini, T. Z., Drabkin, H. J. y Howe, D. (2010). A MOD(ern) perspective on literature curation. *Molecular genetics and genomics : MGG*, 283 (5), p.415–425. [En línea]. Disponible en: doi:10.1007/s00438-010-0525-8 [Consultada: 31 July 2014].

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O. y Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, 455 (7209), Nature Publishing Group, p.47–50. [En línea]. Disponible en: doi:10.1038/455047a [Consultada: 24 February 2014].

Llorente Bousquets, J. E. y Michán, L. (2010). Biodiversidad y biología organísmica. *Ludus Vitalis*, XVIII (33), p.313–316.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. y Hung, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. [En línea].Disponible en: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation [Consultada: 29 September 2014].

Michán, L. (2011). Cienciometría, información e informática en ciencias biológicas: enfoque interdisciplinario para estudiar interdisciplinas. *Ludus Vitalis*, XIX (35), p.239–243.

Michán, L. y Llorente Bousquets, J. E. (2003). La taxonomía en México durante el siglo XX. *Publicaciones Especiales del Museo de Zoología*, Número 12. UNAM. p.229.

Michán, L., Macías, L., Alvarez, E., Muñoz, I., Medina, A. E., Montoya, L. y Bernal, A. (2010). *Propuesta de creación y mantenimiento de un repositorio de literatura institucional en la Facultad de Ciencias, UNAM. (en revisión)*. [En línea]. Disponible en: http://repositorio.fciencias.unam.mx:8080/xmlui/handle/11154/141093.

Michán, L. y Morrone, J. J. (2002). Historia de la taxonomía de coleoptera en México durante el siglo XX: una primera aproximación. *Folia Entomológica Mexicana*, 41 (1), p.67–103.

Michán, L. y Ramírez-Álvarez, D. *Repositorio Ciencias* [en línea]. Sistema de Información Ciencias. Facultad de Ciencias, UNAM. México, D. F. 2013. [Fecha de consulta: 5 Diciembre 2014]. Disponible en:http://repositorio.fciencias.unam.mx:8080/xmlui/handle/11154/139820

Morales Morejón, M., Carrodeguas Rodríguez, M. E. y Avilés Merens, R. (2004). Las intranets en la gestión informacional: un escalón imprescindible en la búsqueda del conocimiento organizacional. *ACIMED*, 12 (3), 2000, Editorial Ciencias Médicas, p.1. [En línea]. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352004000300003&lng=es&nrm=iso&tlng=es [Consultada: 29 September 2014].

NewOCR. (2014). Optical Character Recognition. [En línea] Disponible en: http://www.newocr.com/

RAE. (2014). Diccionario de la Lengua Española. Vigésima segunda edición. [En linea]. Disponible en: http://www.rae.es/recursos/diccionarios/drae

Ramírez Álvarez, D. (2014). *Colección de Literatura de la descripción de nuevas especies para México.* (Seminario de titulación). Facultad de Ciencias. Universidad Nacional Autónoma de México.

Ramírez Martínez, D. (2014). *Sistema de Información Ciencias UNAM: Biología.* (Apoyo a la investigación). Facultad de Ciencias. Universidad Nacional Autónoma de México.

REMIB. (2014). Red Mundial de Información sobre Biodiversidad. CONABIO. [En línea]. Disponible en: http://www.conabio.gob.mx/remib/doctos/remib_esp.html

Rzhetsky, A., Seringhaus, M. y Gerstein, M. (2008). Seeking a new biology through text mining. *Cell*, 134 (1), p.9–13. [En línea]. Disponible en: doi:10.1016/j.cell.2008.06.029 [Consultada: 11 July 2014].

Rzhetsky, A., Shatkay, H. y Wilbur, W. J. (2009). How to get the most out of your curation effort. Bourne, P. E. (ed.). *PLoS computational biology*, 5 (5), Public Library of Science, p.e1000391. [En línea]. Disponible en: doi:10.1371/journal.pcbi.1000391 [Consultada: 31 July 2014].

Salimi, N. y Vita, R. (2006). The biocurator: connecting and enhancing scientific data. McEntyre, J. (ed.). *PLoS computational biology*, 2 (10), Public Library of Science, p.e125. [En línea].Disponible en: doi:10.1371/journal.pcbi.0020125 [Consultada: 25 January 2014].

Sánchez, M., Martínez, A. I. y Alayola, A. (2011). *Informática Biomédica*. México, D.F.: Elsevier, p.177 pp.

St Pierre, S. y McQuilton, P. (2009). Inside FlyBase: biocuration as a career. *Fly*, 3 (1), p.112–114. [En línea]. Disponible en: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2837272&tool=pmcentrez&rendertype=abstract [Consultada: 1 April 2014].

Thessen, A. E. y Patterson, D. J. (2011). Data issues in the life sciences. *ZooKeys*, (150), p.15–51. [En línea]. Disponible en: doi:10.3897/zookeys.150.1766 [Consultada: 2 May 2014].

Wikipedia. (2014). Enciclopedia libre en líjhnea. Disponible en: https://es.wikipedia.org