Acta Botanica Mexicana 108: 81-103 (2014)

LA BIOCURACIÓN EN BIODIVERSIDAD: PROCESO, ACIERTOS, ERRORES, SOLUCIONES Y PERSPECTIVAS

Maribel Castillo[1], Layla Michán[2,4] y Armando Luis Martínez[3]

[1]Comisión Nacional Para el Conocimiento y Uso de la Biodiversidad (CONABIO), Dirección General de Proyectos, 14010 México, D.F., México.
[2]Universidad Nacional Autónoma de México, Facultad de Ciencias, Departamento de Biología Comparada, Laboratorio de Cienciometría, Información e Informática Biológica (CIIB), 04510 México, D.F., México.
[3]Universidad Nacional Autónoma de México, Facultad de Ciencias, Departamento de Biología Evolutiva, Museo de Zoología, 04510 México, D.F., México.
[4]Autor para la correspondencia: laylamichan@ciencias.unam.mx

RESUMEN

La curación de datos biológicos digitales o biocuración es la actividad de organizar, representar y hacer que la información biológica esté accesible para los seres humanos a través de las computadoras. Entre sus tareas están la organización, estandarización, normalización, clasificación, anotación y análisis de la información. El Sistema Nacional de Información sobre Biodiversidad (SNIB) de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) integra la información referente a cerca de seis millones de registros de ejemplares y observaciones biológicas provenientes principalmente de las colecciones zoológicas y herbarios de México. Para administrar esa información la CONABIO ha establecido mecanismos de control de calidad de los datos que ingresan al SNIB que permiten integrar la información proveniente de diferentes fuentes y hacerla consistente e interoperable con otros sistemas de información. Se expone la importancia de la biocuración de bases de datos de biodiversidad, se explica el proceso de curación llevado a cabo en el sistema Biótica© de CONABIO, se dan algunos ejemplos de los errores más comunes que se presentan en los datos biológicos como: omisión, tipográficos, contextuales, redundancia, convención, uniformidad y congruencia; se presentan algunas soluciones, y se discute sobre la importancia de la investigación y enseñanza de la biocuración para los biólogos del siglo XXI.

Palabras clave: bases de datos biológicas, biocuración, biodiversidad, CONABIO, e-taxonomía.

ABSTRACT

The curation of biological digital data or biocuration is the activity of organizing, representing and making biological information accessible to human beings through computers. Among its tasks are the organization, standardization, normalization, classification, annotation and analysis of information. The National System of Information on Biodiversity (SNIB) from the National Commission for Knowledge and Use of Biodiversity (CONABIO) integrates information of about six million records of biological organisms and observations mainly from zoological collections and herbaria in Mexico. To manage this information CONABIO has established quality control mechanisms of the data that are included in the SNIB, that allows to integrate the information of different sources and make it consistent and interoperable with other information systems. This work has the purpose of exposing the importance of biocuration of biodiversity databases, explaining the curating process carried out in Biotica©, which is CONABIO´s Information System, exemplifying some of the most common errors that occur in biological data such as: omission, typographical, contextual, redundancy, convention, uniformity and consistency, presenting some solutions, and discussing the importance of research and teaching of biocuration for biologists of 21st century.

Key words: biocuration, biodiversity, biological databases, CONABIO, e-taxonomy.

INTRODUCCIÓN

Uno de los efectos más evidentes de la (r)evolución digital en la Biología es la creación de enormes volúmenes de datos biológicos primarios en formato digital que se sistematizan en numerosas bases de datos (Schadt et al., 2010; Trelles et al., 2011). El archivo, curación, conservación digital, análisis e interpretación de todos estos datos biológicos son un desafío (Goble et al., 2008). Una disciplina emergente e interdisciplinaria, la informática biológica, produce teorías, métodos y herramientas de vanguardia para lograr este objetivo (Heidorn, 2003). En este documento nos interesa especialmente la curación o mejor dicho, la biocuración, definida como la actividad de organizar, representar y hacer que la información biológica esté accesible para los seres humanos y las computadoras (Howe et al., 2008). Entre sus tareas están la organización, estandarización, normalización, clasificación, anotación y análisis de la información. La mayor cantidad de biocuradores realizan su trabajo en el área biomédica (Burge et al., 2012).

Esta tarea ha sido fundamental para el desarrollo de la biología desde Linneo hasta el GenBank® y el UniProt; la cuidadosa recopilación y organización de los datos biológicos son la base del conocimiento biológico actual. Los nuevos avances tecnológicos y la evolución hacia la web semántica han convertido el proceso de curación en un área emergente, innovadora y de vanguardia (Thornton, 2009). Es tal la importancia de este procedimiento, que hay revistas especializadas sobre el tema como son la DATABASE The Journal of Biological Databases and Curation (http://database.oxfordjournals.org/) y organizaciones que atienden este tema como la International Society for Biocuration (ISB) (http://www.biocurator.org/) y ELIXIR unites Europe's leading life science organisations (http://www.elixir-europe.org/).

Un ejemplo de este tipo de procedimiento y enfoque es el utilizado por la CONABIO para el Sistema Nacional de Información sobre Biodiversidad (SNIB) que integra la información referente a cerca de seis millones de registros curatoriales, bases de datos de tipo taxonómico, ecológico, cartográfico, bibliográfico, etnobiológico, de uso y catálogos sobre recursos naturales y otros temas. Para administrar esa información, la CONABIO ha establecido mecanismos de control de calidad de los datos de ejemplares que ingresan al SNIB que permiten integrar la información proveniente de diferentes fuentes y hacerla consistente e interoperable con otros sistemas de información (Abbott, 2009). Con base en los objetivos de la CONABIO y a la necesidad de hacer eficaz y eficiente el uso y manejo de la información biológica se creó el Sistema de Información Biótica©, diseñado exprofeso para el manejo de datos curatoriales, nomenclaturales, geográficos, bibliográficos y de parámetros ecológicos. Biótica© fue desarrollado tomando en cuenta la gran diversidad de requerimientos de sus principales usuarios, la comunidad biológica (taxónomos, curadores, biogeógrafos, ecólogos, etnobiólogos, entre otros), con el propósito fundamental de ayudar de una forma confiable y sencilla en la captura, actualización y manejo de los datos. En la figura 1 se esquematizan los grupos de información que integran la versión Biótica© 5.0 (CONABIO, 2008).

La versión actual de Biótica© 5.0 está organizada por diez módulos: Base de datos, Directorio, Nomenclatural, Ejemplar, Ecología, Geográfico, Bibliografía, Herramientas y Ayuda, así como un módulo Colecciones que solo será visible si se ha instalado el sistema completo; esto dependerá de las necesidades de captura del usuario**.**

En el módulo Base de datos, se realiza la conexión del sistema a la base de datos donde se ingresa la información; se pueden configurar o predeterminar la visualización específica de los datos que van a utilizarse con frecuencia, lo cual permitirá hacer más rápido y fácil el ingreso de información. También aquí es posible dar de alta a los usuarios y asignarles permisos de acceso a la base de datos.

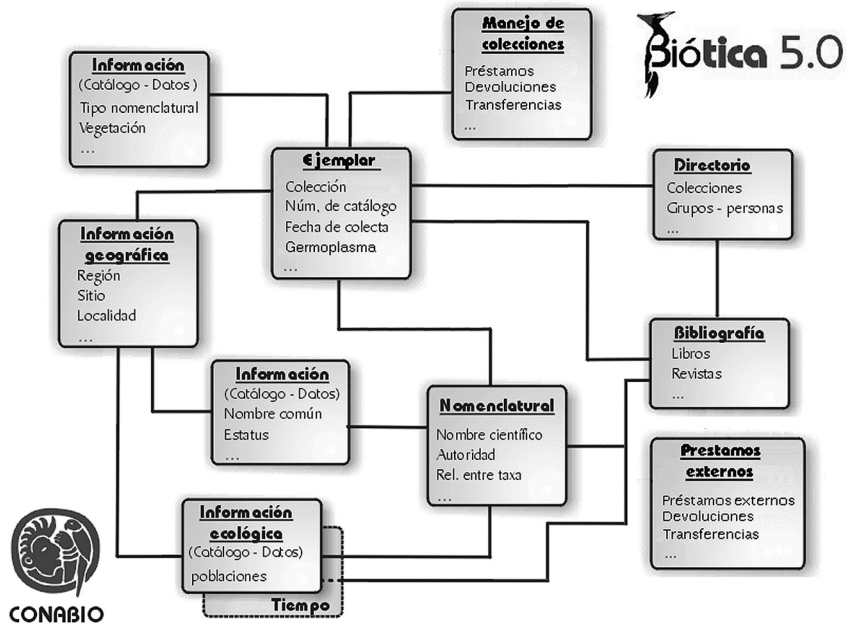

Fig. 1. Esquema general del modelo de datos de Biótica© 5.0.

En el Directorio se captura información de las instituciones y colecciones que resguardan a los ejemplares biológicos como el nombre, siglas, dirección, entre otros, así como grupos de determinadores, colectores u observadores y autores de publicaciones.

La captura y actualización de los nombres científicos con su correspondiente categoría taxonómica, así como los nombres sinónimos, basónimos, etc., se realiza en el módulo Nomenclatural. En esta sección se puede asociar a los nombres científicos, archivos externos al sistema, como imágenes, sonidos, páginas electrónicas, hojas de cálculo; así como, asociar nombres comunes al taxon, a regiones, a características físicas y ambientales y citas bibliográficas.

En el módulo Ejemplar se ingresa la información de la recolecta, observación o registro bibliográfico del ejemplar, como son el nombre científico, la colección a la cual pertenece, sus datos geográficos, el hábitat y microhábitat, las personas que lo colectaron y determinaron, así como la historia de las determinaciones del ejemplar. En esta sección también se puede capturar información biótica y abiótica organizada en su mayoría en catálogos, además es posible la

asociación del ejemplar con archivos externos como imágenes, sonidos, páginas web, hojas de cálculo, etc.

Biótica© contiene un módulo de Información Ecológica, que está dividido en catálogos de parámetros asociados a la población (p. ej. demografía, conducta, reproducción, aspectos físicos del ambiente, etc.) donde es posible clasificar al organismo asociado al estudio (organismo vivo modificado, silvestre, etc.); catálogo de investigadores, que permite ingresar los nombres de los especialistas que llevan a cabo el estudio y poblaciones por taxon donde es posible ingresar datos de una población para toda el área de distribución, o bien para regiones definidas dentro del área de distribución, para todo el periodo de estudio o para una fecha determinada.

En la sección de Información Geográfica se pueden ingresar datos referentes a la localización de los lugares de observación, reporte o recolecta de un ejemplar como son regiones, sitios (coordenadas geográficas o métricas que representan puntos, líneas, polígonos o puntos radio) y localidades. Estos datos pueden estar asociados a la distribución de taxones (regiones), a los nombres comunes y al estudio poblacional. También es posible definir la distribución de taxones mediante la asociación con objetos geográficos (líneas, polígonos y puntos) de mapas digitalizados.

En el módulo Bibliografía se ingresan los datos de las citas bibliográficas (libros, memorias, tesis, artículos, capítulos de libros, entre otros) que pueden relacionarse al ejemplar, al nombre científico o común, a la sinonimia o basónimo, etc., a los catálogos para la nomenclatura, a la información del módulo ecológico y a los préstamos de ejemplares.

Biótica© cuenta con una herramienta de ayuda para el Manejo de Colecciones en lo que se refiere al préstamo, devolución y envío de ejemplares a otras instituciones, así como extensiones de tiempo de los préstamos, envíos, transferencias y devoluciones totales o parciales de material prestado.

La información que se ingresó a la base de datos de Biótica© se puede consultar por medio de una herramienta de reportes dinámicos que puede diseñar el usuario de acuerdo con sus necesidades de información.

La principal fuente de información de la CONABIO sobre la biodiversidad son los proyectos realizados por instituciones de investigación y enseñanza superior del país y organizaciones no gubernamentales; otra fuente es la obtención de datos de especímenes mexicanos que se encuentran en herbarios y museos de otros países. La información de los especímenes que recibe la CONABIO se encuentra en diferentes formatos (Biótica©, entidad-relación, tabla plana, hojas de cálculo, etc.).

Debido a la procedencia de los datos y sus diferentes formatos, es necesario revisar y estandarizar la información de cada una de las bases para que sean compa-

tibles con el SNIB. Para revisar la información de las bases de datos, la CONABIO ha implementado un procedimiento de detección de errores e inconsistencias que permite evaluar la calidad de sus datos, en términos de confiabilidad y exactitud, tanto en los aspectos biológicos, como técnicos. Este procedimiento es conocido como el control de calidad de las bases de datos taxonómicas-biogeográficas. En la figura 2 se muestra de manera general el procedimiento seguido desde que la base de datos de cada proyecto llega a la CONABIO, hasta que se termina la revisión de la información contenida en la base. En el Cuadro 1 se encuentran las herramientas que se utilizaron para realizar el control de calidad.

Cuadro 1. Principales herramientas utilizadas en el procedimiento de revisión, análisis y validación de los datos taxonómicos-biogeográficos para la detección de posibles errores en CONABIO.

Microsoft Access (en varias de sus versiones)	Access, es un software desarrollado por Microsoft®, permite manejar los datos mediante tablas (formadas por filas o registros y columnas o variables), crear relaciones entre tablas, elaborar consultas y formularios para introducir datos o informes para extraer la información.
Automatización de consultas en Access	Una consulta es una pregunta que se hace a la base de datos a partir de requerimientos específicos de información, el resultado de una consulta nos ayuda a recobrar la información para poder analizarla. Las consultas son diseñadas exprofeso para la búsqueda de cada tipo de error y facilitan la revisión de las bases de datos.
Catálogos de autoridades taxonómicas	Son bases de datos de catálogos basados en sistemas de clasificación y arreglos taxonómicos recientes y ampliamente usados por la comunidad científica. Son utilizados en la revisión de los datos de nomenclatura. Algunos disponibles en: http://www.conabio.gob.mx/informacion/catalogo_autoridades/doctos/electronicas.html
Sistema de información Biótica©	Permite el manejo eficiente de datos curatoriales, de nomenclatura, geográficos, bibliográficos y de parámetros ecológicos. Disponible en: http://www.conabio.gob.mx/biotica5/
Verificador de modelos de datos	Es una herramienta que permite comparar estructuras de bases de datos creadas en Access, compara tablas, campos, tipos de datos, campos obligatorios, relación entre tablas, índices, reglas de validación, llaves primarias y foráneas, generando reportes de las diferencias existentes. La herramienta está incluida en el Sistema Biótica©

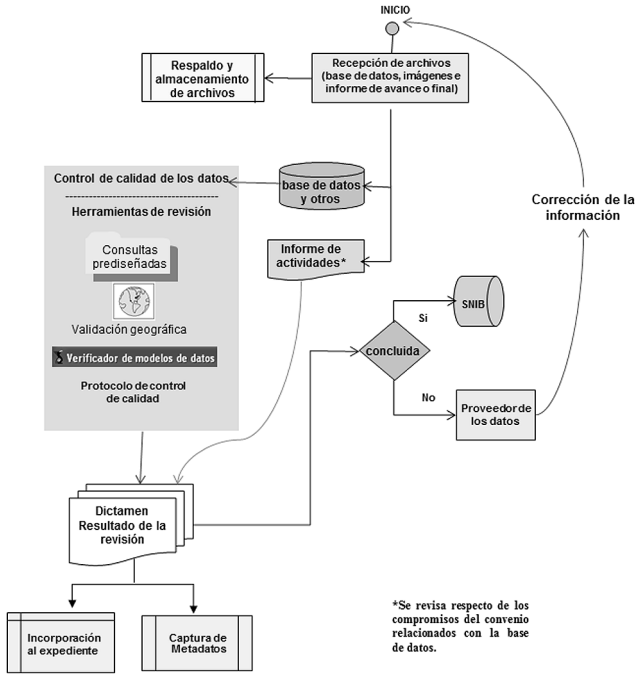

Fig. 2. Diagrama de flujo del procedimiento seguido en la CONABIO desde el ingreso de las bases de datos que aportan datos sobre biodiversidad (modificado de CONABIO, 2006).

MATERIAL Y MÉTODO

El proceso de biocuración se puede clasificar en cuatro etapas: conceptualización, diseño y captura de los datos (en líneas discontinuas), normalización (en líneas continuas), anotación punteada y análisis y publicación en línea heterogénea (Fig. 3).

El proceso de biocuración en la CONABIO inicia con la revisión de la estructura de la base de datos, de los campos y tipos de datos que deben ser acordes con la concepción y planificación de los datos digitales que se van a manejar. Esta revisión es importante ya que en algunas ocasiones el administrador de la base de datos modifica la estructura establecida entre la CONABIO y la institución del proponente, y en consecuencia se altera el contenido. En la mayoría de los casos tales cambios no son realizados de forma adecuada, lo que ocasiona pérdida de información o de integridad de datos; por ejemplo, pueden añadir datos no válidos o modificar los

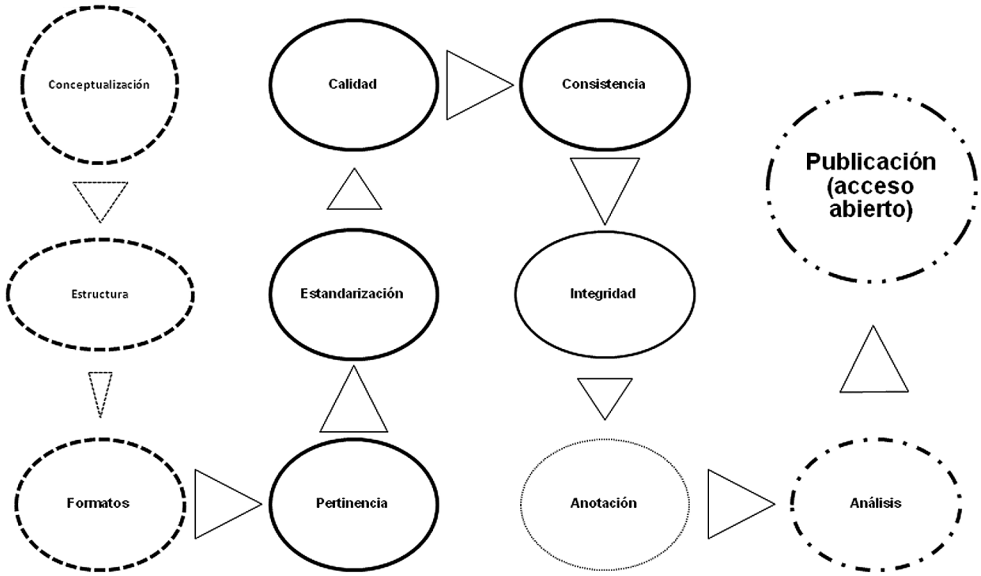

Fig. 3. Etapas y procesos en la biocuración de datos biológicos en formato digital.

existentes tomando un valor incorrecto. También se presentan modificaciones en las bases de datos por errores en el sistema manejador de la base de datos, muchos de ellos ocasionados por fallas en el suministro de energía eléctrica o cambios bruscos de voltaje al momento de guardar la información, o simplemente por medios de almacenamiento defectuosos o contaminados.

La fase denominada normalización establece la calidad de la información de una base de datos y para esto se examina su contenido; para facilitar el manejo y reconocimiento de la información ésta se clasifica en diferentes subconjuntos denominados capas: Curatorial, Taxonómica, Geográfica, Bibliográfica, Colecciones e Instituciones, Colectores y Determinadores. Estos subconjuntos pueden variar según el tipo de proyecto que aporta los datos y se discuten a continuación.

Capa curatorial: Contiene el nombre y siglas de la colección que resguarda al ejemplar, el nombre científico, localidad (sitio de colecta y referencias geográficas del sitio, altitud o profundidad del sitio de colecta), nombre del colector, número de colecta, nombre del determinador, fecha de colecta y determinación (día mes y año), así como información de la restricción de uso de los datos del ejemplar, como son la fecha y motivos de la restricción. Dependiendo del grupo taxonómico, las bases

pueden contener otra información asociada al ejemplar como son: datos merísticos y características morfológicas, entre otros.

Capa taxonómica: Se refiere a los nombres científicos asociados a los ejemplares o involucrados en interacciones biológicas, así como los sinónimos, equivalencias e híbridos. Con ayuda de catálogos de autoridad, se verifica el estatus válido del nombre del taxon, la autoridad y año de la descripción del nombre, se revisa que estén bien escritos y su correspondencia con un sistema de clasificación o listado taxonómico especificado en los acuerdos del convenio con la CONABIO; también se certifica si han sido asociados con un ejemplar u observación de campo.

Capa geográfica: Hace referencia a la información geográfica asociada al ejemplar, como país, estado, municipio y localidad de recolecta. Se comprueba que la información geográfica corresponda con los catálogos del Instituto Nacional de Estadística y Geografía (INEGI). Se revisa que los datos de latitud y longitud estén capturados en el sistema GG:MM:SS (grados, minutos y segundos) y que se indique el método de obtención de la coordenada geográfica. Asimismo se corrobora que las coordenadas geográficas se ubiquen dentro del estado, municipio o región donde se realizó la recolecta, registro visual o evidencia biológica de cada ejemplar.

Capa bibliográfica: Aquí se revisa que las citas bibliográficas tengan los datos mínimos necesarios para localizar una publicación, como son: autores, fecha y título del artículo, capítulo o libro, nombre de la revista, editor/compilador, editorial, ciudad de la publicación, volumen, número y páginas consultadas.

Capa de colecciones e instituciones: Se revisan los datos relacionados con las instituciones, colecciones o herbarios en donde se encuentran resguardados los ejemplares, así como que se haya capturado el nombre y acrónimo oficiales de la colección y de la institución.

Capa de colectores y determinadores: En esta sección se analizan los nombres de los investigadores que colectaron, observaron o determinaron al ejemplar. Se verifica que estén completos con el apellido paterno, apellido materno y el nombre.

La normalización de la información de cada base de datos se realiza utilizando consultas diseñadas exprofeso para cada capa de información y para cada tipo de error y son de varios tipos: selección, actualización de datos, eliminación, creación de tablas y unión de datos, así como consultas más complejas o específicas estandarizadas con SQL (Structured Query Language, por sus siglas en inglés).

Cuando se encuentra información en algunos campos de texto donde no es posible establecer criterios de búsqueda que ayuden a automatizar una consulta, se tiene que hacer una revisión exhaustiva de todos los datos, de manera no automática y con cierto margen de imprecisión o usando una medida meramente estimada. Si

las bases de datos contienen miles de registros, se toma una muestra aleatoria de la información para realizar la revisión, y en cada avance o informe parcial se verifica una nueva muestra de datos hasta haber revisado, de ser posible en toda la información completa.

En la información biológica existen errores que únicamente se pueden detectar, en gran medida, gracias a la formación y experiencia biológica que el revisor de la base de datos tiene sobre el tema y el grupo taxonómico. Un analista con estudios de biólogo tiene los conocimientos suficientes sobre aspectos variados acerca de la biodiversidad, sin embargo, también es muy útil que haya realizado trabajo de campo y de gabinete para tener un panorama más completo sobre el tema.

RESULTADOS

Para cada uno de los tipos de información se identificaron los diferentes errores que frecuentemente cometen los curadores de la base de datos y se clasificaron en siete tipos: omisión, tipográficos, contexto, redundancia, convención, uniformidad y congruencia (Fig. 4).

Para dar una idea acerca del resultado del control de calidad de la información que se efectúa y lo que se puede encontrar en una base de datos, a continuación se muestran algunos ejemplos.

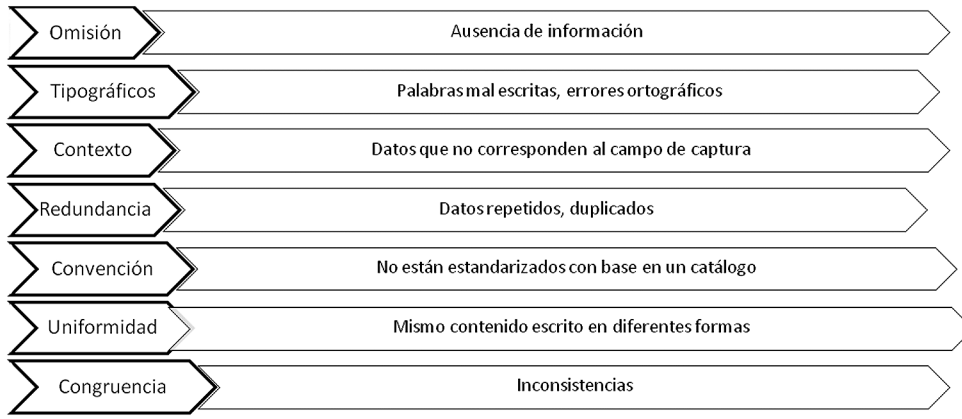

Fig. 4. Los siete tipos de normalización ejemplificados en este trabajo.

Omisión. Se encontraron ejemplares resguardados en una colección entomológica anotados como "sin *Número de catálogo*" (Cuadro 2). Esto parece ser un error ya que los insectos deben de tener número de catálogo cuando están determinados y montados en alfiler entomológico para su ingreso a la colección.

Cuadro 2. Ejemplares con número de catálogo ND = dato no disponible.

Identificador del ejemplar	Categoría taxonómica	Nombre científico	Procedencia del ejemplar	Número de catálogo	Tipo de preparación
1227	especie	*Strymon bazochii*	Colectado	**ND**	Alfiler Entomológico
2765	subespecie	*Baeotis zonata zonata*	Colectado	**ND**	Alfiler Entomológico
2774	subespecie	*Melanis pixe pixe*	Colectado	**ND**	Alfiler Entomológico

Tipográfico. Los errores tipográficos son muy comunes en los campos de tipo texto; por ejemplo, el Cuadro 3 muestra la descripción de localidades que carecen de espacio para separar el texto o bien presentan más de un espacio, asimismo se resalta que hay otro tipo de equivocaciones que se presentan, como la falta el acento en la palabra "Tehuacan" y el uso indistinto de español e inglés en la descripción de una misma localidad.

Cuadro 3. Descripción de localidades con errores tipográficos

Identificador de la localidad	Nombre de la localidad original
338	PUE-OAX, **2 mi S state line**, **Carretera125**
685	**TuxtlaGutiérrez**, 4 km al N.E. de
698	20 km N. **Tehuacan**, cerca de **Loma Colorada**

En el Cuadro 4 se muestran erratas tipográficas que se presentan comúnmente en la información taxonómica, como son: espacios adicionales o falta de espacio en el nombre científico y en el nombre de la autoridad. Asimismo, se resalta un error en la categoría taxonómica donde aparece mal escrita la palabra "especie".

Cuadro 4. Información taxonómica con errores tipográficos.

Identificador del nombre	Categoría taxonómica	Nombre científico	Nombre autoridad
1670	especie	*Heliopsis parvifolia*	**A.Gray**, 1853
1688	especie	*Chaptalia lyratifolia*	**Bur kart**, 1944
1683	**especier**	*Brickellia **lemmo nii***	**A. Gray**, 1882

Contexto. Los errores que se consideran de contexto son datos o parte de los mismos que no corresponden a la definición del campo. Estos datos hay que eliminarlos y capturarlos en el campo que le corresponda. El Cuadro 5 muestra la descripción de una localidad capturada en el campo que corresponde al hábitat y en el Cuadro 6 información que no corresponde al campo *Sexo*, en el que únicamente se captura la condición biológica masculina, femenina o hermafrodita del ejemplar.

Cuadro 5. Información de la localidad capturada en el campo *Hábitat*.

Identificador del ejemplar	Categoría taxonómica	Nombre científico	H**á**bitat
584	subespecie	*Heliconius charithonius vazquezae*	**Tuxtla Gutiérrez, 4 km al NE de**

Cuadro 6. Información que no corresponde al campo *Sexo*

Identificador del ejemplar	Categoría taxonómica	Nombre científico	Sexo
1137	subespecie	*Hemiargus hanno antibubastus*	**Adulto**
1343	subespecie	*Microtia elva elva*	**Colectado**

Redundancia. Si se considera que los números de colecta son únicos por colector y que éstos no se repiten (a menos que el colector realice recolectas por lotes, como en el caso de los peces que se recolectan con redes), son errores de redundancia a los ejemplares del Cuadro 7 que presentan mismos números de colecta, y fueron colecta-

dos por la o las mismas personas. Para descartar un posible desacierto en los ejemplares del Cuadro 7, se debe verificar el método y la fecha de la recolecta del ejemplar, ya sea en la etiqueta del ejemplar o si existiera, en la libreta de campo del colector.

Convención. Son datos capturados sin utilizar convenciones establecidas, ni estándares de información. Por ejemplo, en los nombres de personas los países hispanohablantes utilizan el sistema de doble apellido (paterno y materno). En los países anglosajones y muchos países europeos sólo se utiliza un apellido, normalmente el paterno, precedido de uno o dos nombres.

Cuadro 7. Ejemplares de diferentes especies con mismo número de colecta y mismos colectores.

Identificador del ejemplar	Categoría taxonómica	Nombre científico	Procedencia	Número de colecta	Identificador del grupo	Colectores
584	Especie	*Heliconius charithonius*	Colectado	**10697**	131	Máximo Martínez & Luis Lamberto González Cota
710	subespecie	*Zerene cesonia cesonia*	Colectado	**10697**	131	Máximo Martínez & Luis Lamberto González Cota
1137	subespecie	*Hemiargus hanno antibubastus*	Colectado	**10747**	131	Máximo Martínez & Luis Lamberto González Cota
1343	subespecie	*Microtia elva elva*	Colectado	**10747**	131	Máximo Martínez & Luis Lamberto González Cota
1136	Especie	*Hemiargus hanno*	Colectado	**10748**	131	Máximo Martínez & Luis Lamberto González Cota
1543	subespecie	*Pyrisitia dina westwoodi*	Colectado	**10748**	131	Máximo Martínez & Luis Lamberto González Cota

En el Cuadro 8 se muestran algunos nombres de origen hispano los cuales en el campo Apellido materno presentan una inicial. El acuerdo de captura para esta

información es que en los campos de apellido paterno, materno y nombre de colectores, determinadores y autores de publicaciones, no se debe capturar iniciales. Si no se cuenta con la información completa deberán capturar el dato ND que quiere decir que se trata de información no disponible. Para los nombres que no utilizan el apellido materno, el acuerdo de captura establece que deberán capturar NA que equivale a "No aplica", como se ejemplifica en el registro con identificador de la persona 4724.

Uniformidad. Es importante que exista consenso en la captura del tipo texto, ya que esto facilita la consulta de información. Por ejemplo, se considera un error de uniformidad a los registros del campo "Tipo de preparación" del ejemplar (Cuadro 9), que corresponden a una misma descripción escrita en forma diferente, por lo que en este caso se debe corregir (uniformar) la información a un mismo dato, podría ser: En sobre.

Cuadro 8. Nombres de personas con abreviaturas.

Identificador de la persona	Abreviado	Apellido paterno	Apellido materno	Nombre
4724	M. Douglas	Douglas	**NA**	**ND**
4736	M. Fuentes C.	Fuentes	**C**.	Mario
4690	M. Martínez A.	Martínez	**A.**	Máximo
4596	J. Camelo G.	Camelo	**G.**	Joaquín

Cuadro 9. Datos de tipo de preparación sin uniformidad.

Identificador del ejemplar	Categoría taxonómica	Nombre científico	Número de catálogo	Tipo de preparación
23153	subespecie	*Urbanus dorantes dorantes*	130668	**En sobre**
23126	especie	*Urbanus procne*	130671	**Sobre**
23127	subespecie	*Noctuana lactifera bipuncta*	130672	**En un sobre**
21258	especie	*Urbanus procne*	130672	**En sobres**

Congruencia. Se refiere a datos erróneos, inexactos o inconsistentes. En el Cuadro 10 se muestran taxones que no corresponden a la categoría taxonómica asig-

nada y también la categoría taxonómica a la que realmente corresponden. En el Cuadro 11 se ejemplifica una misma especie descrita por el mismo autor en distinto año.

Cuadro 10. Taxones que no corresponden a la categoría taxonómica asignada.

Identificador del nombre	Categoría taxonómica	Nombre científico	Nombre autoridad	Categoría taxonómica correcta
1534	**subespecie**	*Desmodium psilophyllum*	Schltdl., 1838	**Especie**
1533	**división**	Liliopsida	L., 1753	**Clase**
1529	**subgénero**	*Asclepias*	A. Gray, 1985	**género**

Cuadro 11. Misma especie descrita en diferente año por el mismo autor.

Identificador del nombre	Categoría taxonómica	Nombre científico	Nombre autoridad
625	especie	*Cichlasoma urophthalmus*	Günther, **1862**
169	especie	*Cichlasoma urophthalmus*	Günther, **1864**

Considerando que en México un biólogo termina la licenciatura aproximadamente a los 23 años y suponiendo que a esa edad inicia su labor como colector y que el promedio de esta actividad es de aproximadamente 40 años, un posible error serían los colectores cuyo intervalo de colecta es mayor de 40 años. En el Cuadro 12 se muestran algunos ejemplos en los que al revisar la base de datos se detectó este caso, se solicitó al investigador verificar la información en la etiqueta de los ejemplares asociados a estos ejemplares y se encontró que es correcta, a pesar de que parece incongruente.

En el Cuadro 13 se muestran lugares cuyas coordenadas geográficas fueron verificadas en un mapa, el resultado fue que los sitios se encuentran en un estado diferente al que tienen asociado en la base de datos.

La CONABIO ha documentado el control de calidad de la información de las bases de datos biológicas en un documento de consulta interna llamado “Protocolo

Cuadro 12. Colectores con más de 40 años de trabajo de colecta.

Identificador de la persona	Apellido paterno	Apellido materno	Nombre	Abreviado	Año mínimo	Año máximo	Intervalo
848	Llorente	Bousquets	Jorge Enrique	J. E. Llorente B.	1966	2008	**42**
4452	Pérez	ND	Gonzálo	G. Pérez H.	1945	1994	**49**
3861	Díaz	Francés	Alberto	A. Díaz F.	1937	1996	**59**
412	Escalante	ND	Tarsicio	T. Escalante	1925	1996	**71**

Cuadro 13. Sitios inconsistentes respecto del estado.

Identificador del sitio	Tipo de sitio	Estado según la validación geográfica	Estado según la base de datos	Latitud	Longitud	Resultado
804	Punto	Guerrero	Oaxaca	17° 39' 6"	-99° 50' 24"	Inconsistente
176	Punto	ND	Baja California	28° 37' 3"	-112° 33' 6"	Inconsistente

de control de calidad". Por otro lado, existe un documento llamado "Instructivo para la conformación de bases de datos taxonómico-biogeográficas compatibles con el Sistema Nacional de Información sobre Biodiversidad" (CONABIO, 2010), cuyo propósito es facilitar a los proveedores de las bases de datos biológicas su elaboración, para hacerlos compatibles con el SNIB. En él se describen los campos que son de llenado obligatorio para su ingreso al SNIB que se actualiza cada año y está disponible en la página www.conabio.gob.mx.

DISCUSIÓN

Respecto al manejo de bases de datos biológicas, es importante considerar que desde el inicio de un proyecto se deben diseñar y adecuar los campos a utilizar en las bases de datos que no usan el modelo Biótica©, de tal forma que la información que se repite con mayor frecuencia se capture una sola vez. Biótica© cuenta con catálogos de autoridad precapturados que evitan la duplicidad de información, así como con una sección de configuración en la que el usuario puede predefinir datos

de uso frecuente o repetitivo en el sistema. Por ejemplo, si la base de datos es de un solo estado de la República Mexicana, se define la autocaptura de esa entidad para todos los nuevos registros de la base.

Atomizar datos ayuda a aumentar la calidad de la información; por ejemplo es mejor dividir el nombre científico en dos campos: género y especie, en lugar de mantenerlo unido en un solo. En el modelo de datos Biótica© hay muchos campos atomizados que evitan la captura de información ambigua y permiten agrupar los registros para verificar la información.

Es importante contar con herramientas de verificación de la información tales como gaceteros, que son listados o diccionarios de localidades de una región determinada y pueden contener coordenadas geográficas, nombre del estado, municipio, distrito, catálogos de autoridad taxonómica, listados de nombres, diccionarios de datos, entre otros. La CONABIO en general y el programa Biótica© en particular, cuentan con un número importante de estas herramientas que ayudan a la validación de la información; sin embargo, siempre se debe actualizar, corregir y completar la información existente, es decir debe estar en constante actualización por los expertos de los diferentes grupos taxonómicos a través de proyectos de elaboración y actualización de catálogos.

Considerar que los errores más dificiles de reconocer en una base de datos son los de tipo taxonómico. Estos se pueden originar por una identificación incorrecta o mala determinación del ejemplar. Para evitar esto sería conveniente que se tuviera a disposición ilustraciones o claves ilustradas de las especies que ayuden a verificar la determinación de los ejemplares cuyos datos se van a ingresar a la base, y sería aún mejor contar con el apoyo de un taxónomo especialista y de ser posible corroborar en las colecciones científicas la información antes de capturarla.

Organizar los datos en conjuntos ayuda a la detección de errores, así como en la validación y corrección de la información; por ejemplo el ordenar los datos del mismo tipo con diferentes parámetros como los números de colecta con fecha de colecta y datos del colector permite detectar inconsistencias que no se aprecian si se revisan en campos separados.

La fusión o unión de bases de datos también puede crear nuevos errores, como duplicidad en los registros, por lo que no es recomendable eliminarlos a menos de que se tenga la seguridad de que son duplicados ya que se corre el riesgo de desaparecer información valiosa. Esto aplica cuando un investigador tiene dos bases de datos o más y quiere unirlas para continuar trabajando en una sola.

Es muy importante para el curador tener en todo momento el modelo y las características generales de la estructura y el contenido de cada base de datos (dic-

cionario de campos) para su consulta continua, así como para tener un panorama general de su contenido. También es recomendable documentar la versión del programa con el que fue creada y el manejador utilizado. Estas especificaciones son importantes para alguien que desea utilizarla, ya que podría empezar por consultar los requerimientos del programa que se necesita para poder acceder a la información.

Capacitar a los capturistas en el manejo de la base de datos y proporcionarles, desde el inicio de la captura, estándares *ad hoc* para reducir la tasa de error en la captura y mejorar la calidad de la información.

Para datos de tipo numérico es recomendable guardar las unidades en otro campo, ya que puede haber confusión al momento de capturar el valor numérico y registrarlo con unidades diferentes.

Los biocuradores deben elaborar múltiples respaldos de los datos en diferentes formatos y etapas de la generación de información y almacenarlos en distintos lugares, de preferencia fuera de las instalaciones donde se trabaja. De esta manera se minimiza el riesgo de pérdida de información. La CONABIO cuenta con respaldos en diferentes formatos de almacenamiento de todas las bases de datos, como son cintas magnéticas, discos Blu-ray, DVD y CD; estos se conservan tanto en las instalaciones de la CONABIO como en bóvedas de seguridad en empresas privadas que ofrecen este servicio de resguardo. Los responsables de los proyectos de bases de datos deben considerar la confiabilidad y seguridad del medio de respaldo que elijan; existen nuevas tecnologías que se ofrecen para hacer copias de seguridad, como es la plataforma de usuario conectada a un servidor donde la información está accesible para el usuario en todo momento.

Dado que los registros biológicos son un conjunto de datos complejos cuya interpretación solo es posible si es manejada por especialistas en los diferentes temas que tienen que ver con la biodiversidad, es importante que la gestión de las bases de datos biológicas, su validación, corrección y consulta, sean realizados por biólogos con conocimientos en tecnologías de la información. Es más factible capacitar a un biólogo en las tecnologías de la información y bases de datos que a un especialista en informática en todo el conocimiento biológico que se requiere para interpretar la información biológica.

La normalización determina en gran medida la calidad de la información para que ésta sea consistente, completa, exacta, e íntegra; para lograrlo hay que limpiar y validar los datos implementando estándares de calidad para corregir la información. Cada corrección debe estar documentada y disponible para los usuarios subsecuentes y que ellos sean los que determinen la conveniencia del uso de

los datos, se debe indicar qué controles de calidad se han seguido, qué cambios se han hecho y quién los ha hecho. Esta información debe estar asociada a cada registro correspondiente.

Se puede reducir el esfuerzo y tiempo invertidos en la revisión y validación de la información si se incluye en el diseño de la base algunos campos que indiquen si ha sido verificada (cómo, cuándo, quién los modificó y el resultado de dicha validación). Se puede hacer un archivo externo con esta información asociado a la base de datos, que sería muy útil en los datos de tipo geográfico o en los datos taxonómicos. Esto mejora el proceso del control de calidad y reduce tiempos y costos de revisión, para tener una base confiable y de óptima calidad.

Es importante priorizar la información y el tipo de error a corregir y enfocarse en solicitar únicamente la corrección de los datos que son relevantes, con la finalidad de que los usuarios puedan hacer un uso inmediato de los resultados y no tarde tanto tiempo la corrección de una base de datos. Por ejemplo conviene priorizar la corrección de errores taxonómicos y geográficos por sobre la corrección de texto en los campos de observaciones.

Establecer indicadores de calidad que ayuden a tener una idea clara de qué tan limpios, validados y confiables son los datos de una base; por ejemplo, se puede indicar qué porcentaje de ellos están georreferenciados, el porcentaje de registros geográficos validados in situ, el porcentaje de registros validados por un taxónomo experto, o el porcentaje de datos completos, entre otros. Documentar en un archivo asociado a la base de datos el por qué se capturó un valor "nulo", "no disponible o "no aplica" cuando estos casos se den.

Prevenir los errores es mejor que corregirlos, por lo que hay que asegurar que no vuelvan a ocurrir o que la probabilidad de que sucedan sea menor. Esto se puede trabajar mediante la retroalimentación directa entre los que revisan las bases de datos y los proveedores de la información y viceversa. Para ello es importante la comunicación, misma que se facilita cada vez más por las aportaciones de las nuevas tecnologías de la información y comunicación.

Los avances tecnológicos unidos al desarrollo informático han permitido que los biólogos utilicen estas herramientas, aunadas a sus conocimiento sobre biología, para facilitar el manejo y análisis de gran cantidad de información que obtenemos de la naturaleza, para responder preguntas que ayuden a la toma de decisiones en la administración de recursos naturales, así como para resolver problemas de interés biológico para el bienestar social (Bisby, 2000).

Los biólogos que se han capacitado como técnicos en el manejo de las bases de datos sobre biodiversidad que administra la CONABIO requieren de un perfil

específico común a todos los biocuradores. Preferentemente deben conocer el trabajo que realiza el biólogo en el campo y en las diferentes líneas de investigación (entomología, botánica, mastozoología, herpetología, etc.); también deben tener cierto grado de conocimiento del manejo de colecciones científicas, así como de la historia natural y la distribución geográfica de las especies, además de contar con habilidad para investigar los temas que no aprendieron durante su formación profesional (Sanderson, 2011). Esto permitirá tener una mejor idea de la información que, por ejemplo, debe contener una etiqueta de un ejemplar, dependiendo del taxon bajo estudio o de los datos necesarios para evaluar la situación de los recursos naturales, o especies amenazadas, etc. Asimismo, deben tener la capacidad necesaria para identificar la información que no es propia del taxon, de la colección o del método de colecta, entre otras cosas. La capacitación del personal debe incluir manuales y herramientas computacionales que les sirvan para manejar, organizar, analizar o visualizar información biológica almacenada en las bases de datos (Heidorn et al., 2007).

Sin duda alguna, uno de los requerimientos indispensables es el conocimiento y enseñanza de esta especialidad y enfoques. Los biólogos del siglo XXI deben contar con entrenamiento en el uso de las aplicaciones informáticas para poder incorporarse a este campo laboral actual. De esta forma, es importante que se incluya en los planes de estudio de las carreras biológicas materias relacionadas con la ciencia de la computación (manejo de bases de datos, uso de manejadores de bases de datos como es el Access, Sistemas de Información Geográfica, Percepción Remota, modelado de información, entre otros), como una base más para el manejo de información biológica que ayude a los biólogos egresados a desarrollar proyectos de investigación y a la toma de decisiones para resolver problemas biológicos. Es importante que los planes de estudios fomenten la participación de sus estudiantes en proyectos de investigación, desde su creación (conceptual) hasta las diferentes etapas de desarrollo y su terminación, pero además que participen en la publicación de resultados, todo como parte de su formación profesional. En cuanto a la investigación, se conocen pocos grupos mexicanos interesados formalmente en este tema. Sin duda esta tendencia pronto cambiará y veremos un desarrollo importante en cuanto a la biocuración de las colecciones mexicanas, a través de los productos básicos e indispensables de mayor impacto de esta especialidad, de tal forma que existan colecciones biológicas en línea que preserven, sistematicen y difundan datos biológicos, completos, de calidad, interoperables y de acceso abierto, disponibles para análisis sofisticados entre los que están la cienciometría, la minería de textos y la semántica, que tienen diversas aplicaciones biológicas.

AGRADECIMIENTOS

A dos revisores anónimos que permitieron la mejora sustantiva del artículo. Esta investigación se lleva a cabo gracias al financiamiento del Consejo Nacional de Ciencia y Tecnología, Ciencia Básica, Proyecto 13276 "Análisis de las ciencias biológicas en la actualidad 1980-2010", DGAPA, PAPIME, Proyecto PE212112 "Web 2.0 y 3.0 para dominio de la literatura biológica", PAPIIT IN214212, al IIIC, AC y por su apoyo a las base de datos MARIPOSA y a la CONABIO**.**

LITERATURA CITADA

Abbott, D. 2009. Interoperability. DCC Briefing Papers: introduction to curation. Digital Curation Centre. Edinburgh, UK. Consultado el 31 de marzo de 2014. http://hdl.handle.net/1842/3363

Ball, A. y M. Duke. 2012. How to cite datasets and link to publications. DCC How-to Guides. Edinburg: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/cite-datasets#x1-17000

Bisby, F. A. 2000. The quiet revolution: biodiversity informatics and the internet. Science 289(5488): 2309-2312.

Bourne, P. E. y J. McEntyre. 2006. Biocurators: contributors to the world of science. PLoS Comp. Biol. 2(10): e142. doi:10.1371/journal.pcbi.0020142

Burge, S., T. K. Attwood, A. Bateman, T. Z. Berardini, M. Cherry, C. O'Donovan, C. L. Xenarios y P. Gaudet. 2012. Biocurators and biocuration: surveying the 21st century challenges. Database: The Journal of Biological Databases and Curation. Database (Oxford). 2012: bar059. doi: 10.1093/database/bar059. Consultado el 31 de marzo de 2014. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3308150/

CONABIO. 2002. El Sistema Nacional de Información sobre Biodiversidad. CONABIO. Biodiversitas 44: 3-13.

CONABIO. 2006. Procedimiento para el control de calidad en las bases de datos taxonómicas-biogeográficas que se integran al SNIB. Ver. 1.4. Documento interno, México, D.F., México. 14 pp.

CONABIO. 2008. Sistema de información Biótica©. Versión 5.0. Manual de usuario. México, D.F., México. 977 pp.

CONABIO. 2010. Instructivo para la conformación de bases de datos taxonómico-biogeográficas compatibles con el Sistema Nacional de Información sobre Biodiversidad. México, D.F. México. Consultado el 31 de marzo de 2014. http://www.conabio.gob.mx/web/proyectos/pdf/instructivos/instructivo_bd_2010.pdf

CONABIO. 2012. Sitio oficial de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. http://www.conabio.gob.mx

Davis, A. P., T. C. Wiegers, M. C. Rosenstein, C. G. Murphy y C. J. Mattingly. 2011. The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. Database: The Journal of Biological Databases and Curation 2011(0): bar034. doi:10.1093/database/bar034. Consultado el 31 de marzo de 2014. http://database.oxfordjournals.org/content/2011/bar034.full.pdf+html

EMBL-European Bioinformatics Institute. 2013. ELIXIR European Life Sciences Infraestructure for Biological Information ©. http://www.elixir-europe.org

Goble, C., R. Stevens, D. Hull, K. Wolstencroft y R. Lopez. 2008. Data curation + process curation=data integration + science. Briefings in Bioinformatics 9(6): 506-517. Consultado el 31 de marzo de 2014. http://bib.oxfordjournals.org/content/9/6/506.full

Heidorn, P. B. 2003. Biological informatics: a comparison of biodiversity informatics and neuroinformatics. Bull. Am. Soc. Info. Sci. Tech. 30(1): 12-13. doi: 10.1002/bult.298. Consultado el 31 de marzo de 2014. http://www.asis.org/Bulletin/Oct-03/heidorn.html.

Heidorn, P. B., C. L. Palmer, M. H. Cragin, y L. C. Smith. 2007. Data curation education and biological information specialists. DigCCurr2007: An International Symposium in Digital Curation. Paper and Presentation, Chapel Hill, North Carolina, April 18-20. 2007. North Carolina, USA. Consultado el 4 de abril de 2014. https://www.ideals.illinois.edu/handle/2142/2442

Higgins, S. 2008a. The DCC Curation Lifecycle Model. Proceeding of the 8th ACM/IEEE Joint Conference on Digital Libraries. Pittsburgh, USA. p. 453. doi:10.1145/1378889.1378998

Higgins, S. 2008b. The DCC Curation Lifecycle Model. The International Journal of Digital Curation 3(1): 134-140. doi:10.2218/ijdc.v3i1.48. Consultado el 4 de abril de 2014. http://www.ijdc.net/index.php/ijdc/article/view/69/48

Higgins, S. 2009. The DCC Curation Lifecycle Model, TCDL Bulletin of IEEE Technical Committee on Digital Libraries 5(1): n.d. Recuperado Febrero 22, 2013. doi:10.2218/ijdc.v6i2.191. http://www.ieee-tcdl.org/Bulletin/v5n1/Higgins/higgins.html

Higgins, S. 2011. Digital curation: The emergence of a new discipline. The International Journal of Digital Curation 6(2): 78-88. doi:10.2218/ijdc.v6i2.191. Consultado el 4 de abril de 2014. http://www.ijdc.net/index.php/ijdc/article/view/184

Hirschman, L., G. A. P. Burns, M. Krallinger y C. Arighi. 2012. Text mining for the biocuration workflow. Database: The Journal of Biological Databases and Curation. Database (Oxford). 2012(0): bas020 doi:10.1093/database/bas020. Published online April 18, 2012. http://database.oxfordjournals.org/content/2012/bas020.full

Howe, D., M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. L., Schaeffer, S. St. Pierre, S. Twigger, O. White y S. Y. Rhee. 2008. Big data: the future of biocuration. Nature 455(7209): 47-50. doi: 10.1038/455047a. Published online 3 September 2008. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819144/

Landsman, D., R. Gentleman, J. Kelso y B. F. Francis Ouellette. 2009. DATABASE: a new forum for biological databases and curation. Database (Oxford). 2009(0), bap002–bap002. Published online Mar 16, 2009. doi:10.1093/database/bap002. http://database.oxfordjournals.org/content/2009/bap002.full

Miotto, O., T. W.W. Tan y V. Brusic. 2005. Supporting the curation of biological databases with reusable text mining. Genome informatics 16(2): 32-44. Consultado el 9 de abril de 2014. http://www.jsbi.org/pdfs/journal1/GIW05/GIW05F011.pdf

Salimi, N. y R. Vita. 2006. The biocurator: connecting and enhancing scientific data. PLoS Comp. Biol. 2(10): e1 25. Published online Oct 27, 2006. doi: 10.1371/journal.pcbi.0020125. http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.0020125

Sanderson, K. 2011. BIOINFORMATICS: curation generation. Nature 470: 295-296.

Sarukhán, J. 1992. La coordinación de acciones en torno a la biodiversidad en México: una propuesta de prioridad nacional. In: Sarukhán, J. y R. Dirzo (eds.). México ante los retos de la biodiversidad. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO). México, D.F., México. pp. 291-299.

Schadt, E., M. D. Linderman, J. Sorenson, J. Lee y G. P. Nolan. 2010. Computational solutions to large-scale data management and analysis. Nat. Rev. Genet. 11(9): 647-657.

Shimoyama, M., G. T. Hayman, S. J. F. Laulederkind, R. Nigam, T. F. Lowry, V. Petri, y H. J. Jacob. 2009. The rat genome database curators: who, what, where, why. PLoS Comp. Biol. 5(11): e1000582. doi:10.1371/journal.pcbi.1000582.

Thornton, J. 2009. Data curation in biology – past, present and future. Nature Precedings. doi:10.1038/npre.2009.3225.1.http://precedings.nature.com/documents/3225/version/1

Trelles, O., P. Prins, M. Snir y Jansen, R.C. 2011. Big data, but are we ready? Nat. Rev. Genet. 12(3): 224.

Recibido en octubre de 2013.

Aceptado en abril de 2014.