Manual de Control de Calidad de Datos de Colecciones de Biodiversidad

Dirección General de Repositorios Universitarios

COORDINACIÓN
Tila María Pérez Ortiz
Rubén Sáenz González

CONTENIDOS
Rubén Sáenz González
José Gilberto Parra Leyva
Eréndira González Linares

CUIDADO DE LA EDICIÓN
Mary Carmen Alva Pazarán

Primera edición, octubre de 2023

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Ciudad Universitaria, alcaldía Coyoacán,
C.P. 04510, México, Ciudad de México

DIRECCIÓN GENERAL DE REPOSITORIOS UNIVERSITARIOS
Planta baja, módulo C, Instituto de Biología, 3er. Circuito exterior s/n, Ciudad Universitaria, alcaldía de Coyoacán, C. P. 04510, México, Ciudad de México.
https://dgru.unam.mx
contacto@dgru.unam.mx

Forma sugerida de citar: DGRU (2023). *Manual de control de calidad de datos de colecciones de biodiversidad*. SDI-UNAM. México.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Dr. Enrique Luis Graue Wiechers
Rector

Dr. Leonardo Lomelí Vanegas
Secretario General

Mtro. Hugo Concha Cantú
Abogado General

Dr. Luis Álvarez Icaza Longoria
Secretario Administrativo

Dra. Patricia Dolores Dávila Aranda
Secretaria de Desarrollo Institucional

Lic. Raúl Arcenio Aguilar Tamayo
Secretario de Prevención, Atención y Seguridad Universitaria

Mtro. Néstor Martínez Cristo
Director General de Comunicación Social

Dra. Tila María Pérez Ortiz
Directora General de Repositorios Universitarios

Manual de Control de Calidad de Datos de Colecciones de Biodiversidad

Dirección General de Repositorios Universitarios

Primera edición
Universidad Nacional Autónoma de México
2023

Secretaría de Desarrollo Institucional

DGRU UNAM
Dirección General de Repositorios Universitarios

ÍNDICE

INTRODUCCIÓN

México es un país megadiverso, ya que cuenta con una gran riqueza biológica que se refleja en su amplia variedad de ecosistemas, especies y hábitats. El estudio, la documentación y la difusión de tal diversidad se realizan en las colecciones biológicas, que son esenciales para la conservación, pues preservan la filogenia, distribución, ecología y demás características de las especies que constituyen nuestro patrimonio natural.

La Universidad Nacional Autónoma de México (UNAM) posee un conjunto de colecciones biológicas nacionales que representan una parte significativa de la diversidad biológica del país. Estas colecciones engloban especímenes de plantas, animales, hongos, microorganismos, entre otros elementos. En este contexto, es fundamental gestionar, organizar y resguardar la información generada por estas colecciones biológicas a través de metodologías y tecnologías especializadas. Esto es crucial para lograr diversos objetivos, como asegurar la correcta administración de datos y mantener un riguroso control de su calidad, con el propósito de garantizar la integridad y fiabilidad de la información, la cual servirá como fuente primaria para futuros estudios, investigaciones, labores docentes y otras actividades académicas y científicas.

Desde 2013, la Dirección General de Repositorios Universitarios (DGRU), anteriormente conocida como Coordinación de Colecciones Universitarias Digitales (CCUD), ha desempeñado un papel fundamental en los procesos de estructuración y en el control de calidad de los datos pertenecientes a las colecciones biológicas en la UNAM. Estos procesos y control se ha logrado mediante la implementación de metodologías y herramientas que permiten el registro de información siguiendo estándares internacionales, tales como Darwin Core, junto con la utilización de catálogos reconocidos a nivel global, como World Flora Online, TROPICOS, IPNI, WoRMS, AlgaeBase, entre otros.

Uno de los logros destacados de esta labor ha sido la integración de una gran cantidad de colecciones en un repositorio especializado de colecciones universitarias digitales, denominado Portal de Datos Abiertos UNAM, Colecciones Universitarias, disponible en www.datosabiertos.unam.mx.

Con la experiencia adquirida, la DGRU ha considerado importante plasmar la metodología desarrollada hasta la fecha en este manual. El propósito del presente escrito es proporcionar una lectura básica destinada al personal técnico que se inicia en los procesos de resguardo, administración y control de la calidad de bases de datos relacionadas con colecciones biológicas. Asimismo, está diseñado para quienes deseen familiarizarse con los conceptos básicos, las actividades implicadas y el perfil requerido para llevar a cabo estas tareas.

Este manual se estructura en tres partes principales. Comienza con la presentación de la importancia del control de calidad de los datos, continúa con la descripción de los componentes de la metodología para llevar a cabo este control y, por último, aborda las

metodologías específicas agrupadas de manera semántica en torno a datos biológicos. Estas metodologías se centran en aspectos taxonómicos, temporales y de personal que se relacionan con los eventos de colecta e identificación de organismos, así como en aspectos de ubicación que abordan cuestiones geográficas, coordenadas y localidades.

Se espera que, al final, el lector haya adquirido un conocimiento sólido acerca de los conceptos más relevantes relacionados con el cuidado de la calidad de los datos en colecciones biológicas y esté preparado para llevar a cabo estas tareas de manera efectiva.

ANTECEDENTES

IMPORTANCIA DEL CONTROL DE CALIDAD DE DATOS

> Los costos de hacer inferencias científicas incorrectas basadas en datos erróneos pueden ser sustanciales y de gran alcance: los errores pueden ser sutiles, las conclusiones inapropiadas pueden permanecer sin discutir durante años en la literatura y las investigaciones posteriores pueden verse seriamente comprometidas (Radziwill, 2006).[1]

La sistematización de la información en bases de datos permite almacenarla de forma estructurada con accesos eficientes. Se pueden automatizar procesos y llevar a cabo la integración con sistemas y aplicaciones para aprovecharla de mejor manera. Además, las bases de datos son el punto de partida ideal para aplicar métodos y mecanismos que brindan seguridad a la información y permiten compartirla y analizarla. Esto se logra mediante el uso de tecnologías de la información y métodos computacionales que se han desarrollado y madurado desde hace décadas.

Sin embargo, esta sistematización no exime la necesidad de poder confiar en la información almacenada en las bases de datos. Si bien otorga herramientas para alcanzar dicha confianza, es necesario llevar a cabo procedimientos para controlar la calidad de los datos almacenados desde un punto de vista cualitativo, lo que implica tomar en cuenta relaciones con otras fuentes de datos y aplicar procesos que permitan identificar inconsistencias y errores dentro de un contexto específico, así como la aplicación de procedimientos de limpieza y validación de información.

Existen diversos problemas inherentes a la administración de las bases de datos, los cuales van desde aspectos internos, como diseños ineficientes, redundancia y fragmentación de datos hasta aspectos externos, como el manejo inadecuado de usuarios y privilegios, estrategias ineficaces de escalamiento y vulnerabilidades de seguridad. También están los problemas relacionados con la exactitud de los datos, la inconsistencia, la omisión de información o la falta de validación con otras fuentes de información, errores en la captura de los datos, entre otros. Estos problemas afectan la calidad de los datos, un tema que abordará este manual.

La información de los ejemplares resguardados en las colecciones biológicas se ha sistematizado en bases de datos desde principios de siglo, las cuales han desempeñado un papel esencial en la gestión y accesibilidad de estos datos para ser usados como fuente primaria para diversas actividades académicas, incluida la investigación y colaboración científica. Por esto, cobra particular relevancia controlar la calidad de los datos almacenados a fin de mitigar los efectos adversos de los errores que puedan presentar. En la experiencia

[1] Traducción propia.

de la DGRU, los errores más comunes que se presentan en bases de datos de colecciones biológicas y que afectan la calidad de los datos son:

Datos incompletos: Falta de datos o datos parciales, que pueden derivar en un panorama incompleto y llevar a conclusiones incorrectas. Un ejemplo de esto es la falta de información temporal con respecto a eventos de colecta.

Datos inconsistentes: Cuando los datos se comparan con catálogos de autoridad, pueden presentar diferencias o variantes, lo que puede conducir a una posible inconsistencia, haciendo que sean imprecisos o incorrectos. Por ejemplo, los nombres de las especies en la base de datos de una colección pueden diferir de los listados en catálogos de autoridad taxonómica.

Datos obsoletos: Si los datos no se actualizan periódicamente, pueden quedar obsoletos y llevar a decisiones mal informadas.

Errores de captura: Son aquellos errores que las personas generan en el momento de registrar los datos, como errores tipográficos, valores incorrectos o interpretaciones equivocadas. Este es uno de los errores más comunes en las bases de datos.

Falta de validación: Se refiere a la falta de procesos de supervisión o validación de los datos para ubicar errores e inconsistencias.

Omisión en el manejo de la privacidad de los datos: Comúnmente, las bases de datos almacenan información considerada sensible, como datos personales, valor de bienes o su ubicación. En el caso de las colecciones biológicas, existen especies en peligro de extinción, cuya información de su ubicación reside en bases de datos, y esta debe ser cuidadosamente resguardada y no difundida para evitar la explotación de los ejemplares aún libres en la naturaleza.

Falta de documentación y gobierno de datos: La falta de documentación de la metodología, estándares y políticas utilizadas para el manejo de los datos.

El control de calidad de datos en el ámbito de la gestión de bases de datos, específicamente en colecciones biológicas, es necesario para contar con datos confiables y precisos, lo que resulta fundamental para potenciar la investigación científica. Esto permite a biólogos y otros especialistas llevar a cabo análisis de patrones de distribución, evolución y variabilidad de especies de manera eficiente, así como documentar nuevas especies, dar seguimiento de poblaciones en peligro y definir áreas prioritarias para la conservación. Además, el control de calidad de datos desempeña un papel crucial en la toma de decisiones relacionadas con la conservación y el manejo de recursos naturales. Al poner estos datos a disposición del público y de la comunidad científica, las bases de datos de colecciones biológicas fomentan la transparencia, la colaboración, la adopción de decisiones fundamentadas y la mejora de la calidad de la investigación. En última instancia,

contribuyen al avance del conocimiento en el campo de la biología y la conservación de la biodiversidad.

COMPONENTES DEL CONTROL DE CALIDAD DE DATOS

A continuación, se definirán y describirán las partes y actores que intervienen en un proceso de control de calidad de datos, algunos de ellos ya mencionados. Se asume que el lector posee conocimientos en el manejo de bases de datos y en la gestión de colecciones biológicas y su sistematización.

Base de datos

Bertone definió bases de datos como "colección o conjunto de datos interrelacionados con un propósito específico vinculado a la resolución de un problema del mundo real" (2011).

En el contexto de este manual, una base de datos de una colección biológica se refiere al conjunto de datos de cada uno de los ejemplares preservados en la colección. Estos datos incluyen información taxonómica, geográfica, del evento de colecta y de las personas involucradas en el proceso curatorial del espécimen. A esta delimitación se le denomina dominio o contexto de la información, el cual es la pauta para relacionar los datos entre sí y con otras fuentes de información, como catálogos y vocabularios controlados.

En la actualidad, cuando se habla de bases de datos, también se refiere a los sistemas de gestión de bases de datos (SGBD), que son un conjunto de programas informáticos que permiten consultar, manipular y administrar el acceso a la base de datos. Existen diversos SGBD con diferentes paradigmas, licencias de uso y capacidades. El SGBD utilizado en la DGRU y en las operaciones descritas en este manual es PostgreSQL, que es un sistema de código abierto con una sólida reputación de fiabilidad, flexibilidad, documentación y soporte para estándares técnicos abiertos. PostgreSQL tiene sus antecedentes en desarrollos que se remontan a 1986 e incorporó el soporte para el lenguaje de consulta estructurado (SQL, por sus siglas en inglés) que le dio su nombre actual (IBM, s.f.). Es importante resaltar que PostgreSQL es capaz de gestionar bases de datos relacionales, contando con la implementación de diversos algoritmos y funciones que hace eficiente el acceso y manejo de datos, asimismo provee una escalabilidad adecuada para manejar grandes volúmenes de datos con una alta demanda.

Una base de datos expresa y almacena los datos en las siguientes estructuras:

Tablas: Son el componente principal de las bases de datos relacionales. Son una estructura que consta de renglones y columnas, cada renglón representa un registro único de datos, y cada columna representa un campo de datos específico y de la misma naturaleza.

Registros: Representan las características de un elemento u objeto único, como datos de una persona, de un ejemplar biológico o de las relaciones entre ellos.

Campos: Son características únicas de un registro y se representan en columnas de la tabla. Los campos de una columna deben tener las mismas características estructurales o tipo de datos, y pueden agruparse a lo largo de los registros, lo que permite realizar filtrados, consultas y comparaciones entre registros de la misma tabla o con otras tablas.

Estándares de datos

La información almacenada en una base de datos debe ser correctamente modelada para conseguir una estructura que resuelva o mitigue los problemas relacionados con las bases de datos y el dominio de aplicación. Además, debe otorgar una representación adecuada de los datos para su uso extensivo y la generación de nueva información a partir de ella. En términos de la interoperabilidad de los datos, es necesario considerar seguir un estándar de datos, como Darwin Core, para compartirlos con otras personas u organizaciones.

Para el manejo de datos de colecciones biológicas, se recomienda utilizar el estándar Darwin Core (DwC), que "ofrece un marco de trabajo estable, sencillo y flexible para recopilar datos de biodiversidad provenientes de fuentes diferentes y variables. Darwin Core fue desarrollado originalmente por la comunidad de Biodiversity Information Standards (antes TDWG)" (GBIF, s.f.).

Los campos o términos, como los denomina DwC, están organizados en nueve categorías o clases:

Figura 1. Categorías del Núcleo de Darwin

Record-level Terms	Dublin Core terms, institutions, collections, nature of data record	Simple Darwin Core (flat)
Occurrence	evidence of species in nature, observers, behavior, associated media, references.	
Event	sampling protocols and methods, date, time, field notes	
Location	geography, locality descriptions, spatial data	
Identification	linkage between Taxon and Occurrence	
Taxon	scientific names, vernacular names, names usages, taxon concepts, and the relationships between them	
GeologicalContext	geologic time, chrono-stratigraphy, biostratigraphy, lithostratigraphy	
ResourceRelationship	explicit relationships between identified resources (e.g., one organism to another, taxon to location, etc.)	Generic Darwin Core (relational)
MeasurementOrFact	measurements, facts, characteristics, assertions, references	

Fuente: Wieczorek *et al.*, 2012, "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard".

Las categorías principales se refieren al evento de colecta del ejemplar biológico, a la ubicación del evento, su contexto geológico, ocurrencia, taxonomía e identificación. El resto de las categorías sirven para registrar relaciones, mediciones y más información sobre el registro. Se puede consultar el listado completo de términos en http://rs.tdwg.org/dwc/terms/ y en el *Estándar para datos de biodiversidad Darwin Core* de la DGRU, disponible en https://dgru.unam.mx/wp-content/uploads/2019/10/D.ST_.DGRU_CDI_007_2015_E_Datos_Biodiversidad_Darwin_Core.pdf.

Catálogos

Los catálogos, que incluyen listados de especies y de localidades, desempeñan un papel esencial como fuentes de información fundamental para el estudio no solo de la biología, sino que también para otras áreas como la planificación territorial y la inversión gubernamental (Clasificación de Especies Colombia, s.f.). Además, se utilizan como material de apoyo en la docencia para difundir la riqueza biológica entre la población, concienciar sobre su importancia y promover la conservación de la naturaleza. En términos del control de calidad de datos, los catálogos funcionan como listas de autoridad que permiten contrastar y validar información almacenada en la base de datos. Por ejemplo, es posible validar que los nombres de especies en la base de datos sean consistentes con catálogos taxonómicos para identificar posibles errores de registro, validez de los nombres y otros problemas estructurales de la información. Además, los catálogos pueden poner de manifiesto problemas semánticos que requieren la revisión de curadores de colecciones, investigadores o académicos. Estos problemas pueden tratarse de especies endémicas que no coinciden con la ubicación que se reporta, especies cuyo nombre registrado no concuerda con las características que le acompañan, distribuciones extraordinarias, entre otros.

En este manual se utilizarán, fundamentalmente, dos tipos de catálogos de referencia:

Catálogos taxonómicos: Catálogos que listan nombres aceptados de las especies, así como sus características y distribución.

Catálogos geográficos: Catálogos oficiales que listan los nombres de localidades en una región geográfica, de acuerdo con la división política. También pueden incluir información de la delimitación territorial a través de una descripción en polígonos para sistemas de información geográfica (SIG).

Una ventaja adicional del uso de catálogos es que confieren la capacidad de relacionar la información de varias bases de datos a través de la interoperabilidad semántica. Este principio básico establece que, si se utilizan los mismos identificadores en diferentes acervos para un elemento común, es posible vincular y enriquecer la información. Por ejemplo, cuando se colecta un ejemplar de cierta especie y se identifica utilizando un nombre aceptado de un catálogo, es posible recuperar y conocer información adicional del ejemplar

de otras bases de datos, como los listados de especies en peligro de extinción, distribución de la especie en localidades aledañas o en otras regiones del mundo. El vínculo generado mediante la buena práctica de utilizar catálogos conduce a un mayor aprovechamiento del trabajo previo realizado en todo el mundo, registrado en bases de datos.

Más adelante, en este manual, se proporcionará información sobre los catálogos utilizados en la metodología para controlar la calidad de los datos.

Vocabularios controlados

En estrecha relación con los catálogos, existen datos que requieren cierta consistencia y homogeneidad estructural o semántica para ser útiles o para mitigar los posibles problemas de acuerdo con su tipología y que, posiblemente, no estén registrados en catálogos de autoridad. Estos datos requieren lineamientos específicos para su registro y generalmente se utiliza un vocabulario controlado para ello.

> Un vocabulario controlado es una estructura organizada de palabras y frases usadas para indexar contenido y/o para recuperar contenido a través de la navegación o búsqueda. Típicamente, incluye términos preferidos y sus variantes y describe un dominio específico o tiene un alcance específico (COAR, 2018).

El uso de un vocabulario controlado o restringido ofrece distintas ventajas:

Estandarización: Se establece el uso de términos comunes de forma homogénea o estandarizada, lo que reduce la ambigüedad e incrementa la facilidad de verificación de la información.

Interoperabilidad: Facilita la compartición de datos entre diferentes sistemas y acervos que usen los mismos términos, impulsando así el aprovechamiento de la información entre los diferentes acervos y generando posibles descubrimientos.

Consistencia y calidad: Al tener términos de referencia bien establecidos, es posible identificar errores o inconsistencias en los datos, pues cualquier registro que difiera del estándar puede ser rápidamente identificable.

Análisis de la información: El uso de términos estandarizados a nivel estructural y semántico deriva en que el análisis de la información pueda realizarse de manera más eficiente, pues es posible utilizar herramientas de cómputo para aplicar diversas operaciones con la información.

Gestión a largo plazo: Cuando se establece una política de uso de vocabularios controlados, la gestión de la base de datos tiende a permanecer homogénea a lo largo del tiempo y al cambio del personal que la realiza, pues hay un marco de referencia que seguir y se minimiza el uso de criterios personales para el registro de información.

En el caso de las bases de datos de colecciones biológicas, se recomienda utilizar vocabularios controlados en campos donde se registren nombres de especies, ubicaciones geográficas, tipos de hábitats, estados de vida, tipo de ejemplares, coberturas temporales, unidades de medida, marcadores genéticos, categorías y todos aquellos que sean susceptibles de utilizarse para referir a otros dominios, acervos, o bien, para hacer análisis de la información a través de operaciones, filtros, agrupaciones, etcétera.

Herramientas de datos

La administración de las bases de datos de colecciones biológicas y la aplicación de los conceptos descritos en las secciones anteriores no serían posibles, al menos no de una manera eficiente, sin el uso de sistemas que nos ofrezcan herramientas para tratar, comparar, transformar, analizar o compartir los datos.

A fin de inducir a que el lector tome la decisión más adecuada a sus circunstancias, en este manual no se hará referencia a operaciones utilizando herramientas específicas, más allá del SGBD que utiliza la DGRU. A continuación, se deja una lista de herramientas que son útiles para diferentes escenarios, dependiendo del volumen de datos, las operaciones a realizar, la arquitectura de sistemas considerada, el nivel de análisis necesario y la preferencia de herramientas de código abierto o propietarias:

PostgreSQL: Sistema de gestión de bases de datos relacional de código abierto, que se destaca por su robustez, extensibilidad y conformidad con los estándares SQL. Es conocido por ser altamente escalable y adecuado para aplicaciones empresariales y de gran volumen.

MySQL: Sistema de gestión de bases de datos relacional de código abierto, conocido por su velocidad y eficiencia, popular para aplicaciones web. Es fácil de instalar y ofrece diversas funcionalidades fundamentales como la integridad transaccional.

Apache Cassandra: Base de datos distribuida, no relacional diseñada para manejar grandes volúmenes de datos distribuidos en varios servidores, otorgando alta disponibilidad y tolerancia a fallas.

Apache Spark: Ambiente de trabajo para el procesamiento de datos en tiempo real y por lotes, utilizado para análisis de datos a gran escala.

OpenRefine: Herramienta de código abierto para la limpieza y transformación de datos. Proporciona funciones para explorar, transformar y limpiar datos mediante algoritmos de forma automatizada.

Microsoft Excel: Parte de un paquete de herramientas de ofimática, Excel permite analizar y gestionar un conjunto relativamente limitado de datos, con una estructura principalmente uniforme expresada en una tabla.

RapidMiner: Plataforma de ciencia de datos que provee una amplia gama de herramientas para la preparación de datos, análisis predictivos, etc. Soporta diversos algoritmos para el análisis y modelado de datos.

KNIME: Herramienta de código abierto que sirve para análisis de datos, reportes e integración. Permite construir procesos de tratamiento de datos para hacer análisis de datos.

Tableau: Herramienta de visualización de datos que permite crear gráficos interactivos y paneles de control para el análisis de datos. Es ampliamente utilizado en empresas para la toma de decisiones basadas en datos.

Python + bibliotecas: Lenguaje de programación popular para el análisis de datos, que, utilizando bibliotecas como Pandas, NumPy y Matplotlib, facilita la manipulación y visualización de datos.

R: Lenguaje de programación utilizado en estadísticas y análisis de datos, pues ofrece una amplia variedad de paquetes y herramientas específicas en estos temas.

PowerBI: Herramienta de visualización de datos producida por Microsoft®, que es compatible con otras herramientas de la marca, como Excel y Azure.

Pentaho: Paquete de herramientas de código abierto para la gestión y análisis de datos. Incluye componentes para la extracción, transformación e inserción, así como herramientas de creación de informes y análisis.

Hadoop: Plataforma de código abierto que se utiliza para el procesamiento y análisis de grandes conjuntos de datos. Incluye el sistema de archivos distribuidos y el entorno de procesamiento MapReduce.

Personal

Es imperativo que el personal responsable de controlar la calidad de los datos cuente con un perfil adecuado para realizar las actividades y manejar las herramientas mencionadas, debido a la trascendencia de la información en el momento de generar nuevo conocimiento. El personal que controle la calidad de los datos debe contar con habilidades técnicas y habilidades blandas que le permitan, por un lado, realizar procedimientos y operaciones en los datos y, por otro lado, lidiar con los problemas inherentes al proceso y encontrar la forma de resolverlos para obtener los resultados esperados. Lo anterior, debe estar alineado a la tecnología y procedimientos que se decida usar para el control de calidad y, en el caso de este manual, es necesario saber conceptos básicos en el manejo de datos de colecciones biológicas. A continuación, se detallan algunas habilidades básicas para el personal que realiza el control de calidad:

Conocimientos en programación: Muchas de las operaciones de carga y transformación de datos requerirán la programación de procedimientos automatizados, por lo que es ideal que el personal maneje el lenguaje de programación seleccionado para estos fines.

Uso de herramientas de datos: El análisis y limpieza de datos requerirán que el personal conozca cómo utilizar herramientas que le permitan realizar su trabajo de manera más precisa, eficiente y sencilla.

Administración de bases de datos: Para el resguardo y transferencia de datos es fundamental que el personal conozca el manejo de un sistema de gestión de bases de datos.

Conocimientos en el manejo de datos biológicos: Los términos, catálogos y vocabularios controlados requieren que el personal comprenda cómo funcionan las colecciones científicas y cómo identifican y preservan sus ejemplares.

Resolución de problemas: Es importante que el personal pueda abordar problemas, idear soluciones o descomponerlos en problemas más sencillos.

Pensamiento crítico: La habilidad para evaluar datos y resultados de forma crítica, identificando posibles problemas o sesgos en el análisis.

Investigación e iniciativa: No existe un manual o capacitación que provea de los conocimientos para resolver todos los aspectos y problemáticas relacionados con el control de calidad, por lo que es fundamental que el personal sea capaz de investigar, aprender y aplicar nuevas técnicas o herramientas en su trabajo.

Comunicación: El personal debe poder comunicar y explicar de forma precisa y detallada los resultados y hallazgos a compañeros y al personal no técnico.

Trabajo en equipo: Dada la diversidad de datos o su volumen, podría ser necesario que un equipo de analistas trabaje de forma coordinada y eficiente, lo que requiere la capacidad de trabajar en equipo.

En suma, el personal que realiza control de calidad requiere un perfil especializado que debe formarse en aspectos técnicos, tecnológicos y de gestión.

Evaluación de la calidad de datos

Antes de adentrarnos en las metodologías de control de calidad de los datos en colecciones biológicas, es importante destacar que la evaluación de la calidad de los datos es fundamental para garantizar su confiabilidad y veracidad. Esta evaluación implica la implementación de mecanismos que permitan medir la calidad de los datos, lo que a su vez facilita la toma de decisiones sobre su uso.

En este manual, se presentará la metodología utilizada por la DGRU para determinar la calidad de diferentes tipos de datos y la técnica que utiliza para expresarla, la cual se basa

en la asignación de calificaciones o indicadores de calidad de acuerdo con criterios para cada campo. Además, se mostrarán los criterios para determinar el nivel de calidad tomando como base su calificación. Esto permite saber el estado general de un conjunto de datos de forma ágil y eficiente.

A continuación, se presentan algunas de las evaluaciones que la DGRU realiza a las bases de datos de las colecciones biológicas, las cuales están publicadas en el *Manual de Datos Abierto de Colecciones Universitarias Digitales* (2017, pp. 80-81):

a) Si el dato existe.
b) Si es consistente de origen: existe en catálogos y es coherente en su relación con otros campos.
c) Si está desactualizado o no tiene lógica por sí mismo.
d) Si está desactualizado o no tiene lógica en el conjunto.
e) Si se aplica una modificación que no afecta su significado (acentos, mayúsculas, traducción de un nombre, otros).

El objetivo de esta evaluación permite:

a) Detectar datos anómalos.
b) Reducir la presencia de errores.
c) Proveer datos homogéneos.
d) Ubicar datos duplicados, datos faltantes e incompletos, errores ortográficos,
e) datos ambiguos, datos desactualizados y datos obsoletos.
f) Contar con datos de calidad.

Derivado de este análisis, los datos se clasifican en tres tipos:

a) Consistentes con los catálogos. Se consideran correctos y de calidad.
b) Modificados en el proceso. Son consistentes con catálogos, pero requieren de una validación por parte del curador o responsable de los datos.
c) Inconsistentes con los catálogos: se reportan al curador o responsable de la colección para su revisión.

METODOLOGÍAS ESPECÍFICAS PARA EL CONTROL DE CALIDAD DE LOS DATOS

Las metodologías específicas desarrolladas en este manual se agrupan semánticamente en varios dominios: datos taxonómicos, geográficos (incluyendo coordenadas y localidades), nombres de personas y fechas (para colecta e identificación) que convergen dentro del evento de colecta. Esta clasificación atiende a la necesidad de revisar las categorías en las que con mayor frecuencia existe información para los ejemplares biológicos, como la identificación taxonómica, ubicación geográfica y temporal, así como quién realiza la colecta u observación del ejemplar.

Taxonomía

Dentro del ámbito de la información relacionada con los ejemplares de colecciones biológicas, la identificación taxonómica de los registros de ejemplares es uno de los principales dominios de datos que se incluyen en las bases de datos. Esta información puede ser registrada en el momento de la colecta o actualizada posteriormente en uno o varios eventos de identificación, con el propósito de mantener siempre la identificación lo más completa y actualizada posible.

Esta identificación es fundamental para distinguir un registro de otro cuya información se complementa con los datos del evento de colecta (geográfico y temporal), entre otras observaciones. Esto, a su vez, determina los datos principales que conforman un registro de la presencia de un taxón específico en un lugar determinado del mundo dentro de la base de datos.

El control de calidad en los datos taxonómicos está enfocado principalmente en la estandarización de la escritura de los nombres científicos que tienen los registros de ejemplares. Esta estandarización se lleva a cabo con el respaldo de catálogos especializados confiables y actualizados, a fin de reducir las variantes de escritura que puede haber para los registros de ejemplares que pertenecen a un mismo taxón. Esto, a su vez, optimiza la precisión en las búsquedas de información dentro de la base de datos y reduce la presencia de nombres incorrectos, duplicados, anómalos o desactualizados (DGRU, 2022).

En consecuencia, la elección de estándares, catálogos, así como el diseño de las bases de datos que albergan esta información, son cruciales para un buen control de calidad.

Estándares de datos

El uso de estándares especializados en datos de naturaleza biológica es una práctica ampliamente adoptada por instituciones y organizaciones que se dedican a la gestión de

datos de colecciones biológicas (Ortega y Guevara, 2017). Emplear un estándar como modelo para la creación de la base de datos necesaria es un buen punto de partida.

El estándar más ampliamente usado para datos biológicos es Darwin Core (Darwin Core Task Group, 2009) (DGRU, 2019). Este estándar incluye una sección denominada "Taxón" que abarca toda la clasificación taxonómica, junto con una serie de campos relacionados con las referencias del nombre científico, como el estatus taxonómico, la fuente del nombre, los nombres aceptados, el código de nomenclatura empleado, entre otros. Por lo tanto, establecer clases semánticas de datos dentro del estándar (Wieczorek *et al*., 2012) permite identificar los datos originales de manera que converjan en un sistema de clasificación común, independiente de su fuente.

Base de datos

El diseño de la base de datos curatorial debe basarse preferentemente en un estándar ampliamente aceptado, como Darwin Core (Darwin Core Task Group, 2009), esto permite tanto el mantenimiento de la base a través de la curación de los datos existentes como su ampliación mediante la integración de nuevos registros.

Los datos que se integran en la base de datos suelen proceder de distintas fuentes, como etiquetas, registros de libros, frascos e incluso bases de datos previas, que pueden tener diferentes formatos e incluso estar desarrollados en distintos programas informáticos. Sin embargo, el uso del estándar facilita el desarrollo de un criterio de ingreso y codificación similar sin importar la fuente de los datos de los registros. Es decir, no importa si se trata de datos de registros de ejemplares botánicos, zoológicos o bacterias; los campos del estándar están diseñados para poder incluir todos los datos del ejemplar de una forma lógica y coherente abarcando categorías específicas (Wieczorek *et al*., 2012) lo que permite recopilar de manera efectiva todos los datos que corresponden a los ejemplares.

Considerando los campos principales de la clase "Taxón" e "Identificación" de Darwin Core (Darwin Core Maintenance Group, 2021), algunos de los campos son obligatorios en el control de calidad mientras que otros son opcionales y dependen en gran medida del nivel de identificación del ejemplar y de los catálogos usados como referencia. A continuación, se presenta una lista de los campos principales de las clases "Taxón" e "Identificación" del estándar de datos biológicos Darwin Core que suelen integrarse durante el control de calidad de datos taxonómicos.

Cuadro 1. Campos de las clases "Taxón" e "Identificación" del estándar Darwin Core empleados en el control de calidad de datos taxonómicos

Tipo campo	*Nombre campo Darwin Core*	*Estatus en control de calidad*
Campo de registro[a]	higherClassification	Opcional
	kingdom	Opcional
	phylum	Opcional

Tipo campo	*Nombre campo Darwin Core*	*Estatus en control de calidad*
	class	Opcional
	order	Opcional
	family	Obligatorio
	genus	Opcional
	subgenus	Opcional
	specificEpithet	Opcional
	infraspecificEpithet	Opcional
	taxonRank	Obligatorio
	verbatimTaxonRank	Obligatorio
	scientificNameAuthorship	Opcional
	scientificName	Obligatorio
	acceptedNameUsage	Opcional
	nameAccordingTo	Opcional
	nomenclaturalCode	Opcional
	taxonomicStatus	Opcional
	nomenclaturalStatus	Opcional
	taxonRemarks	Opcional
	identificationRemarks	Opcional
	identificationQualifier	Opcional
Campos de control	lastModified	Obligatorio
	lastModifiedUser	Obligatorio
Identificadores	occurrenceID	Obligatorio
	uuid	Obligatorio
	datasetID	Obligatorio

[a] Para todos los "campos de registro" en la base de datos, se incluyen campos con el mismo nombre que el campo de registro, seguido de un sufijo para indicar los valores originales (sufijo "_o"), las calificaciones (con sufijo "_qi") y los permisos o banderas (con sufijo "_pub"). El nivel mínimo de determinación que debe tener un registro es a nivel de familia. El número de campos opcionales que se completan para las categorías taxonómicas dependerá del nivel de identificación taxonómica del registro, por ejemplo, el campo "infraspecificEpithet" suele permanecer vacío en registros determinados a nivel de especie.

Fuente: Elaboración propia con base en la información de *Darwin Core Quick Reference Guide* (Grupo de Mantenimiento Darwin Core, 2021).

Catálogos

Un elemento esencial en el control de calidad de los datos taxonómicos es la selección adecuada de los catálogos taxonómicos que se emplean para estandarizar y validar los nombres científicos de los registros (Torres-Mejía *et al.*, 2016). Esta elección tiene como finalidad seleccionar los catálogos más completos y actualizados periódicamente, ya que los nombres científicos están sujetos a cambios conforme se realizan nuevas investigaciones y revisiones taxonómicas.

En general, son los proveedores de datos de las colecciones quienes determinan qué catálogos son apropiados para utilizar durante el proceso de control de calidad. Algunas de las cualidades que deben presentar estos catálogos incluyen que contengan la información definida como obligatoria en Darwin Core, con los nombres científicos enlistados (atomizados o concatenados) que contengan su autoridad y año de publicación, su rango taxonómico, un identificador único dentro del catálogo para poder asignar al campo *nameAccordingTo*, así como el estatus taxonómico y el nombre aceptado en caso de ser un nombre sinónimo o inválido.

En caso de que el catálogo no tenga definidos los mismos nombres de campos que establece Darwin Core, se debe estandarizar los nombres de los campos al integrar el catálogo taxonómico.

Respecto a la clasificación superior, es común que muchos catálogos no la incluyan, es decir, no contengan categorías taxonómicas por encima de familia, como reino, *phylum* o división, clase y orden. Sin embargo, pueden asignarse con base en la familia o el género utilizando la clasificación de los proveedores de datos.

Es necesario considerar que los catálogos pueden ser generales o específicos para un grupo biológico en particular. Por ejemplo, existen catálogos específicos para grupos como plantas, como The Plant List (The Plant List, 2013), recientemente actualizado en World Flora Online (WFO, 2023), la base de datos del Missouri Botanical Garden: TROPICOS (Tropicos, 2023a) y el International Plant Names Index (IPNI, 2023). Otros catálogos están especializados en grupos que comparten características biológicas en común, como AlgaeBase (Guiry y Guiry, 2023), enfocado en algas y organismos afines, como diatomeas y dinoflagelados, o el World Register of Marine Species (WoRMS Editorial Board, 2023) enfocado en especies marinas sin importar el grupo biológico al que pertenezcan los organismos.

Por otro lado, existen catálogos más amplios que contienen información sobre nombres científicos descritos en casi todos los reinos taxonómicos, como el Integrated Taxonomic Information System (National Museum of Natural History & Smithsonian Institution, 2023), Catalogue of Life (Bánki *et al*., 2023) o portales agregadores como GBIF (GBIF, 2023).

Algunos catálogos están enfocados a la biota de una región o país en específico, como el desarrollado por la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), en su sistema de Catálogos de Autoridades Taxonómicas (CAT) (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad [CONABIO], 2023), el cual incluye un listado de las especies presentes en México junto con su clasificación taxonómica.

En última instancia, dado las numerosas opciones de catálogos disponibles, son los proveedores de datos de las colecciones quienes determinan el orden de prioridad de uso, de modo que, si las opciones de búsqueda se agotan en el primer catálogo, se puede continuar al que sigue en orden de prioridad hasta completar la revisión de los nombres científicos.

Vocabularios controlados

En el caso de los datos taxonómicos, generalmente se emplean las recomendaciones del estándar Darwin Core (Darwin Core Maintenance Group, 2021) para ciertos campos específicos, ya que las categorías taxonómicas ya se encuentran estandarizadas por medio del uso de los distintos catálogos. Algunos de los campos que emplean vocabularios controlados son:

- ***taxonRank***: incluye nombres generalizados para especificar el rango taxonómico más específico de un nombre científico, como orden, familia, género, subgénero, especies, subespecie, variedad, forma, entre otros.
- ***taxonomicStatus*****:** incluye la estandarización de distintas variantes para indicar el mismo estatus según el código de nomenclatura, como Aceptado y Sinónimo para plantas, algas y hongos, y Válido e Inválido para animales.
- ***nomenclaturalStatus*****:** incluye la estandarización de distintas abreviaturas del estatus nomenclatural del nombre del taxón, según las reglas de conformación de nombres de cada código de nomenclatura, como Nombre conservado, Nombre legítimo, Nombre ilegítimo, Nombre dudoso, Nombre rechazado, entre otros.

Herramientas de datos

Las herramientas básicas que se emplean en la comparación con los catálogos taxonómicos corresponden con consultas de comparación y actualización. Estas pueden realizarse en cualquier gestor de base de datos o programas similares que permitan comparar los datos taxonómicos de origen con los catálogos de referencia. Con dichas consultas se puede crear una "tabla de trabajo" la cual se emplea para comparar su información contra los catálogos taxonómicos.

El proceso de control de calidad de los datos taxonómicos comprende siete fases principales que abarca la búsqueda, revisión, comparación, asignación y validación de la información proveniente de los catálogos. A continuación, se describen estas fases y las actividades clave que se desarrollan en cada una de ellas.

1. Establecimiento de jerarquía de los catálogos
    a. Es importante establecer antes de la revisión el orden de importancia de los catálogos de consulta. Esto normalmente es decidido por el proveedor de datos y determina a cuál catálogo se le da la prioridad para validar los nombres científicos.
    b. Por lo general, deben agotarse las opciones de similitud con el primer catálogo para el total de los registros antes de comenzar con el segundo catálogo en orden de prioridad. Esto se realiza con la finalidad de no crear heterogeneidad causada por artefactos de la revisión dentro de los catálogos, ya que, aunque los catálogos pueden contener los mismos taxones, al tener una referencia distinta, crea variantes en el proceso de asignación y validación de datos.

2. Determinación de rangos taxonómicos presentes
    a. Identificar los diferentes rangos taxonómicos de los registros de ejemplares en la base de datos es una buena práctica para determinar cuáles de los campos obligatorios y opcionales del estándar para datos taxonómicos deben ser contemplados para su revisión (Cuadro 1).
3. Selección de jerarquía de rangos taxonómicos para la revisión
    a. Se recomienda revisar un rango taxonómico a la vez. Por ejemplo, si se elige el rango de especie, se deben revisar todos los nombres con ese rango en todos los catálogos antes de pasar a otro rango.
    b. Se recomienda comenzar con el rango de especie, ya que es el más común en los registros de ejemplares. Luego, se pueden revisar los rangos de infraespecíficos (por ejemplo, subespecie, variedad, forma) o género. Después de revisar estos rangos, se recomienda revisar los rangos superiores a género, como familia, orden, clase, etc. Estos niveles de determinación taxonómica no son muy comunes, ya que en la mayoría de las bases de datos en los que se han realizado procesos previos de control de calidad taxonómica se encuentran revisados a niveles más específicos.
    c. Después de revisar cada rango, es común que algunos registros no puedan ser validados directamente con los catálogos taxonómicos, por lo que estos registros se marcarán como inconsistentes para ser revisados por el proveedor de datos.
4. Comparación y asignación de los valores de campos atomizados
    a. Una vez seleccionado el rango taxonómico a revisar, se pueden comparar los elementos del nombre científico con los elementos correspondientes presentes en el catálogo taxonómico.
    b. El número de campos que deben coincidir dependerá del rango taxonómico, pero en todos los casos se debe emplear una consulta de comparación.
    c. Las comparaciones se realizan entre los campos originales (marcados con sufijo "_o") y el campo correspondiente en el catálogo taxonómico. Las comparaciones, de manera general, pueden ir desde la igualdad en el valor hasta la similitud en distintos rangos, por lo que se emplean algoritmos de similitud.
    d. El número de comparaciones entre los elementos del nombre científico y sus equivalentes en el catálogo taxonómico variará en función de las diversas formas de escritura que presente el valor original del nombre científico.
        i. La primera comparación se realiza de manera que los nombres coincidan. Por ejemplo, para un taxón a nivel de especie, se deben verificar que los campos *genus*, *specificEpithet* y *scientificNameAuthorship* sean idénticos al catálogo. Para un taxón infraespecífico, se deben comparar los campos *genus*, *specificEpithet*, *infraspecificEpithet*, *scientificNameAuthorship* y el nombre del rango infraespecífico, como variedad, subespecie o forma, para asegurarse de que coincidan.
        ii. Las comparaciones posteriores se realizan por medio de los algoritmos de similitud con los elementos del nombre científico, empezando por los más específicos y avanzando hacia los más generales. Por ejemplo, en el caso de un taxón a nivel de especie, se puede buscar que los campos *genus* y *specificEpithet*

coincidan con el catálogo, mientras que para el campo *scientificNameAuthorship*, se utilizan algoritmos de similitud para determinar la opción más cercana al valor original de *scientificNameAuthorship*. En el caso de un taxón a nivel de infraespecie, se sigue un criterio similar, verificando que los campos *genus*, *specificEpithet* y *infraspecificEpithet* concuerden con el catálogo, y aplicando los algoritmos de similitud al campo *scientificNameAuthorship*.

iii. Estas comparaciones pueden realizarse hasta que se obtenga una única opción dentro de los nombres del catálogo que sea congruente con el valor original del nombre científico.

e. Como se mencionó anteriormente, los campos para revisar y asignar información a partir del catálogo y sus comparaciones dependen del rango taxonómico de los registros de los ejemplares. En el Cuadro 2 se muestran los campos que deben considerarse para los rangos taxonómicos más comunes en los que se encuentran determinados los registros de ejemplares, junto con las posibles variantes en los valores de *taxonRank* tanto en el nombre científico original como en los catálogos.

Cuadro 2. Consultas de comparación empleadas para los elementos del nombre científico según el "Rango taxonómico" del registro del ejemplar

Rango taxonómico	*genus*	*specificEpithet*	*infraspecificEpithet*	*scientificNameAuthorship*	*taxonRank (variantes)*
Infraespecie (subespecie, variedad o forma)	Revisión y comparación	Revisión y comparación	Revisión y comparación	Revisión y comparación	subespecie, ssp., subesp. | variedad, var., varierty | forma, fo., f.
especie	Revisión y comparación	Revisión y comparación	No aplica	Revisión y comparación	especie, sp.
género	Revisión y comparación	No aplica	No aplica	Revisión y comparación	género, gen.

Nota. Se presentan las consultas aplicables a los campos del estándar Darwin Core, como *genus*, *specificEpithet*, *infraspecificEpithet* y *scientificNameAuthorship*, en función del "Rango taxonómico". La columna "*taxonRank* (variantes)" exhibe las distintas formas de escritura correspondientes a los distintos valores de "Rango taxonómico" que pueden encontrarse en los nombres científicos originales o en los catálogos taxonómicos.

Fuente: Elaboración propia a partir de los nombres de los campos de Darwin Core (Grupo de Mantenimiento Darwin Core, 2021).

f. Una vez identificada la versión del nombre científico original dentro del catálogo, se actualiza la "tabla de trabajo" con los datos seleccionados en la consulta de comparación provenientes del catálogo taxonómico. Esta actualización incluye información disponible en el catálogo, en campos como *taxonomicStatus*, *nomenclaturalStatus*, *acceptedNameUsage*, *nameAccordingTo* y *nomenclaturalCode*. Además, se incorporan los datos relacionados con la clasificación superior, que pueden incluir *kingdom*, *phylum*, *class*, *order* y *family*, siempre que estén disponibles en el catálogo.

5. Asignación o revisión de la clasificación superior (taxonomía superior)

a. Cuando la clasificación superior se asigna mediante la consulta de actualización desde el catálogo taxonómico, los campos correspondientes se evalúan de manera similar a los campos del nombre científico, lo que implica comparar los valores originales de los campos "kingdom", "phylum", "class", "order" y "family".
b. En caso de que el catálogo taxonómico no contenga dichos campos, es posible asignarlos a partir de otro catálogo previamente seleccionado para ese fin con la autorización del proveedor de datos. En este escenario, se puede tomar como referencia el nombre del campo "genus" o "family" para realizar la consulta de comparación y actualizar la "tabla de trabajo".

6. Concatenaciones de campos
   a. Algunos campos del estándar Darwin Core corresponden a concatenaciones de campos atomizados, como *scientificName* y *higherClassification*. Estas concatenaciones varían en función del rango taxonómico.
      i. En el caso de *higherClassification* deben concatenarse todos los nombres de los rangos inmediatos superiores al rango taxonómico más específico que tenga el nombre científico. En el Cuadro 3 se muestra un ejemplo de cómo se debe presentar la información concatenada según el estándar Darwin Core.
      ii. En cuanto al nombre científico, la forma de concatenación debe ajustarse al código de nomenclatura correspondiente. Por ejemplo, el año de publicación debe formar parte del campo de para el Código Internacional de Nomenclatura Zoológica (ICZN) (ICZN, 1999), mientras que según el Código Internacional de Nomenclatura para algas, hongos y plantas (ICBN) (Turland *et al.*, 2018) no se integra como parte del campo *scientificName*. Lo mismo ocurre con el rango infraespecífico de subespecie, que debe ser incluido en el campo para organismos como plantas, hongos y algas, según el ICBN como "subsp." (Turland *et al.*, 2018), mientras que, según el ICZN, no se incluye, ya que las subespecies se indican como un nombre trinomial (ICZN, 1999).

Cuadro 3. Ejemplos de concatenación de los campos según el rango taxonómico presente en el nombre científico y el código de nomenclatura correspondiente

Rango taxonómico	*scientificName (Animalia)*	*scientificName (Plantae)*	*higherClassification*
Infraespecie (subespecie)	genus + specificEpithet + infraspecificEpithet + scientificNameAuthorship Acanthogorgia multispina typica Kükenthal & Gorzawsky, 1908	genus + specificEpithet + taxonRank + infraspecificEpithet + scientificNameAuthorship Oxalis articulata subsp. rubra (A. St.-Hil.) Lourteig	kingdom | phylum | class | order | family | genus | specificEpithet
Infraespecie (variedad o forma)	genus + specificEpithet + taxonRank + infraspecificEpithet + scientificNameAuthorship	genus + specificEpithet + taxonRank + infraspecificEpithet + scientificNameAuthorship Oxalis articulata fo. crassipes (Urb.) Lourteig	kingdom | phylum | class | order | family | genus | specificEpithet

Rango taxonómico	*scientificName (Animalia)*	*scientificName (Plantae)*	*higherClassification*
	Acanthogorgia multispina var. iridescens Kükenthal & Gorzawsky, 1908		
especie	genus + specificEpithet + scientificNameAuthorship Acanthogorgia multispina Kükenthal & Gorzawsky, 1908	genus + specificEpithet + scientificNameAuthorship Oxalis articulata Savigny	kingdom | phylum | class | order | family | genus
género	genus + scientificNameAuthorship Acanthogorgia Gray, 1857	genus + scientificNameAuthorship Oxalis L.	kingdom | phylum | class | order | family
familia	family + scientificNameAuthorship Paramuriceidae Bayer, 1956	family + scientificNameAuthorship Oxalidaceae R. Br.	kingdom | phylum | class | order
orden	order + scientificNameAuthorship Malacalcyonacea McFadden, van Ofwegen & Quattrini, 2022	order + scientificNameAuthorship Oxalidales Bercht. & J. Presl	kingdom | phylum | class

Nota. Se aplica el Código Internacional de Nomenclatura Zoológica para el reino Animalia y Código Internacional de Nomenclatura para algas, hongos y plantas para el reino Plantae. Dentro del campo *higherClassification* pueden incluirse otras categorías taxonómicas que pueden estar presentes en el valor original del nombre original y que pueden ser validadas con los catálogos taxonómicos, como *subphylum*, *suborden* y *subclase*.

Fuente: Elaboración propia a partir de los nombres de los campos de Darwin Core (Grupo de Mantenimiento Darwin Core, 2021).

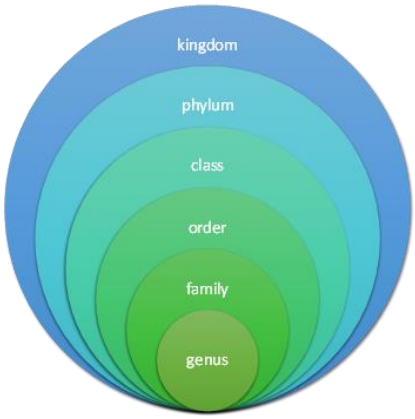
b. La agrupación de las categorías taxonómicas se basa en diferentes códigos de nomenclatura según el reino:
   i. Para el reino Plantae, la agrupación se rige por el "International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code)" (Código Internacional de Nomenclatura para algas, hongos y plantas), que fue adoptado durante el Decimonoveno Congreso Botánico Internacional en Shenzhen, China, en julio de 2017. Esta referencia se encuentra en la publicación "Regnum Vegetabile 159" bajo la edición de Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J., y Smith, G. F. Puede accederse a esta referencia en línea a través de https://doi.org/10.12705/Code.2018.
   ii. En cambio, para el reino Animalia, se toma como base el "International Code of Zoological Nomenclature" (Código Internacional de Nomenclatura Zoológica) en su cuarta edición publicada en 1999. Este código es mantenido por "The International Trust for Zoological Nomenclature" y puede consultarse en detalle

en https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/.

iii. Para ejemplos de nombres científicos en el reino Plantae, se pueden encontrar en el sitio web https://www.tropicos.org/home. Por otro lado, para ejemplos de nombres científicos en el reino Animalia, se pueden explorar en https://www.marinespecies.org/index.php.

7. Revisión y validación de los valores asignados y sus campos de referencia
   a. Después de asignar los valores procedentes del catálogo, se debe revisar y evaluar cada uno de los campos con respecto a los originales (la asignación de las calificaciones se abordará en la sección "Evaluación de la calidad de datos").
   b. Finalmente, es esencial verificar las coherencias de las categorías que presentan anidación (clasificación superior) y de asociación de fuente de referencia (nombre científico).
      i. En el caso de la clasificación superior de categorías taxonómicas, se debe verificar que todos los registros de ejemplares con el mismo valor de género estén asignados dentro de la misma familia y, a su vez, que todos los que estén incluidos en ella pertenezcan a un solo "orden", y así sucesivamente con el resto de las categorías de taxonomía superior, como se ilustra en la Figura 2.

Figura 2. Ejemplo de anidamiento de las categorías taxonómicas presentes en el estándar Darwin Core



Nota. Se muestran los campos del estándar Darwin Core, que incluyen "*kingdom*", *phylum*, *class*, *order*, *family* y *genus*, con su relación anidada.

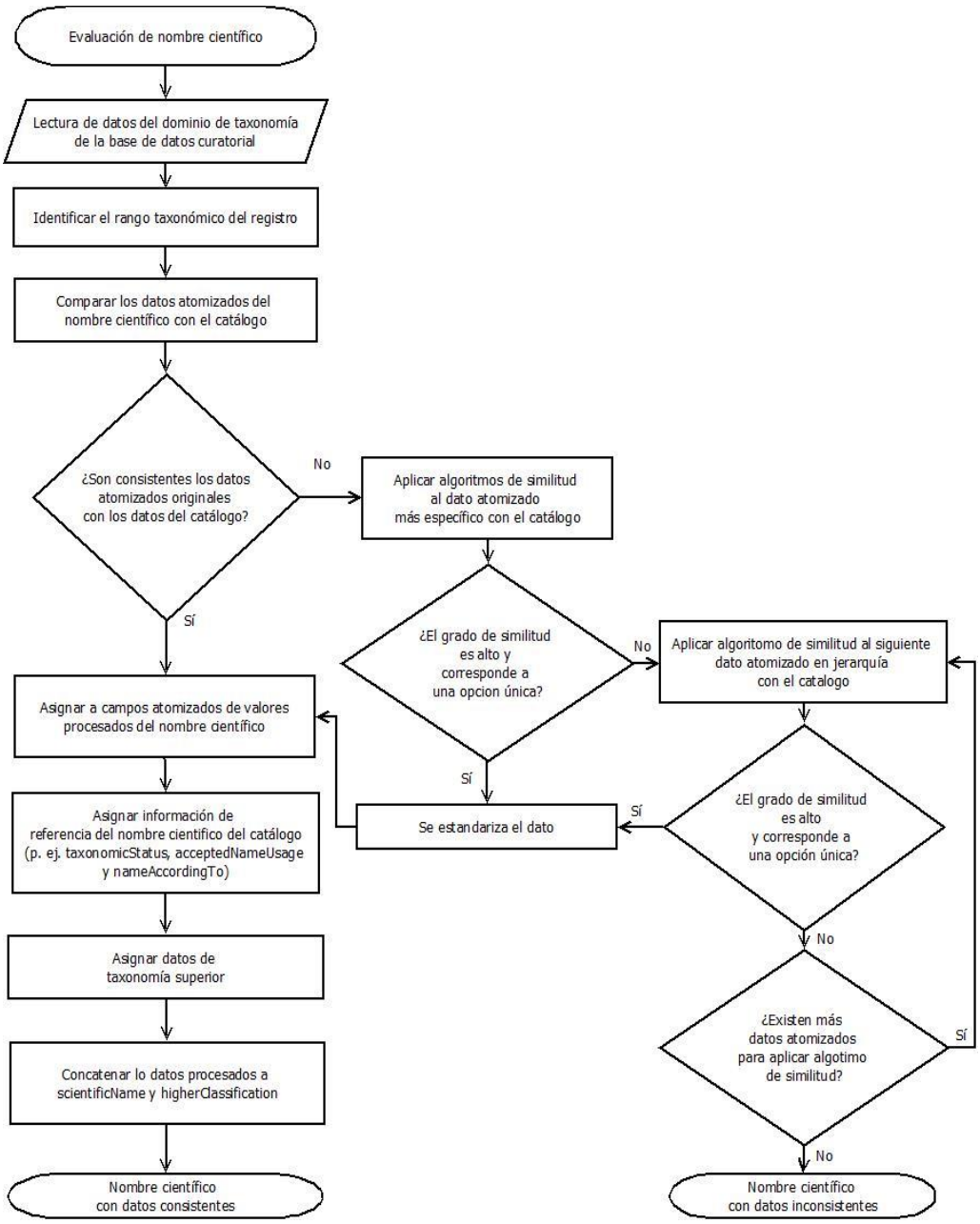
Fuente: Elaboración propia con base en la información del capítulo.

ii. En cuanto a los nombres científicos, debe existir solo una fuente de referencia para la misma combinación de un taxón y su autor (*scientificName*). Por ejemplo, para todos los registros, dentro de la base de datos, nombrados como *Oxalis latifolia* Kunth debe existir solo una fuente de catálogo de referencia (*nameAccordingTo*), al igual que la información de los campos *taxonomicStatus*, *nomenclaturalStatus* y *acceptedNameUsage* (asociados a dicha referencia). Si, al término de la revisión de los nombres científicos, se detectan múltiples fuentes para un mismo nombre, es

imperativo homogeneizar la versión utilizando la fuente del catálogo con mayor prioridad.

Las fases del proceso de control de calidad de datos del dominio de taxonomía se resumen en el diagrama de flujo presente en la Figura 3.

Figura 3. Diagrama de flujo de la revisión de los datos del dominio de taxonomía



Nota. La elección del orden de prioridad de los catálogos se establece previamente en colaboración con el proveedor de datos, por lo que esta metodología se puede aplicar a cualquier catálogo. Este procedimiento aplica para todos los rangos taxonómicos que puedan existir en la base de datos de origen. Los algoritmos de similitud

se pueden aplicar tantas veces como sea necesario, en función del número de elementos que componen el nombre científico, dependiendo de su rango taxonómico.

Fuente: Elaboración propia.

Evaluación de la calidad de datos

La evaluación de la calidad de datos implica la asignación de calificaciones, que resultan de comparar de los valores originales (campos con sufijo "_o") de los campos respecto a los obtenidos con la revisión del catálogo dentro de la "tabla de trabajo" (campos sin sufijo). Para ello se hace uso de los campos de calificación (con sufijo "_qi"), en los que se asigna una de las calificaciones que se mencionan en la sección 3 de este manual.

En términos generales, se califica cada campo de forma individual. En el Cuadro 4 se muestran las calificaciones que son aplicables a todos los campos de taxonomía. Estas calificaciones se dividen en dos grupos principales: consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Cuadro 4. Lista de calificaciones (qi) empleadas en el control de calidad de datos taxonómicos

	Calificación (qi)	*Situación*	*Ejemplos*
consistente	1	Consistente de origen: el dato original y procesado son iguales.	genus_o: Oxalis genus: Oxalis
	0.9	Modificado por variantes de escritura (por ejemplo: digitación, ortografía, cambio de minúsculas o mayúsculas, completar y espacios en blanco). La identidad del dato se mantiene.	genus_o: Oxali**s** genus: Oxali**x**
	0.8	Dato asignado durante el proceso de control de calidad, es decir, que no está presente en el campo original.	kingdom_o: sin dato kingdom: **Plantae**
	0.7	Modificado por estandarización con respecto al catálogo.	Se ha identificado un cambio necesario en la atribución del nombre científico "Oxalis debilis". El autor original, que figura como "H.B.K.", es incorrecto y debe corregirse a "Kunth". La forma correcta de escritura es: scientificNameAuthorship_o: **H.B.K.** scientificNameAuthorship_o: **Kunth**
inconsistente	0.4	Requiere la resolución del proveedor de datos tras realizar la revisión en el control de calidad.	La combinación de epíteto infraespecífico no existe, solo existe una opción posible en el catálogo. Sin embargo, es esencial revisar la propuesta de modificación, ya que esta altera la identidad del taxón en cuestión: scientificName_o: Oxalis hirta var. tubif**olia** scientificName: Oxalis hirta var. tubi**flora** **T.M. Salter**

	0.1	No fue posible establecer una conexión lógica con la definición del campo según el estándar ni se ha podido verificar a través de ningún catálogo o referencia, lo que conduce a considerarlo como inconsistente.	Para el genus *Oxalis* no existe la especie "palmatum", por lo que el valor original no sufre modificación: specificEpithet_o: palmatum specificEpithet: palmatum
nulo	0	El dato original se eliminó del campo.	infraspecificEpithet_o: **L.** infraspecificEpithet: sin dato
	Vacío (sin dato)	Cuando no existe dato para revisar.	Cuando un taxón a nivel de especie y el campo infraspecificEpithet permanece vacío: infraspecificEpithet_o: sin dato infraspecificEpithet: sin dato

Nota. Se muestra la codificación para cada valor de calificación (qi), la situación que describe y algunos ejemplos de los casos en los que se ocupa dicha calificación. Las calificaciones se dividen en tres categorías: consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Fuente: Elaboración propia.

A pesar de que cada campo posee su propio campo original, procesado y de calificación, algunos campos se califican con base en aquellos que los construyen (concatenación). Esto aplica para los campos *higherClassification* y *scientificName*, por lo que para asignarlos se toma en cuenta los elementos que los conforman. Sin embargo, pese a tener en consideración lo anterior para asignar la calificación a los campos concatenados, se debe tomar en cuenta si existe un valor original para dicho campo, en el caso de *higherClassification* y *scientificName*.

Por otro lado, se debe tomar en consideración que, al existir un grado de anidación en las categorías taxonómicas, las calificaciones se pueden heredar en sentido de su consistencia o inconsistencia. Es decir, que para considerar que un nombre científico es consistente, todos sus elementos deben tener una calificación consistente (0.6 o superior). De modo que, un nombre científico no puede ser consistente si su valor para *genus* está calificado como 0.1, a pesar de que el resto de los elementos del nombre sean consistentes, incluida la clasificación superior. Por lo tanto, para un nombre científico inconsistente, la situación de inconsistencia se indicará a partir del rango taxonómico donde el nombre no tenga una referencia en alguno de los catálogos. Por ejemplo, si para un nombre científico no existe la combinación de *genus* + *specificEpithet* + *infraspecificEpithet*, debido a que para esa combinación de nombre (*genus* + *specificEpithet*) no existen infraespecies (*infraspecificEpithet*), el campo de *infraspecificEpithet* será calificado como 0.1, y todos los elementos subsecuentes, como *scientificNameAuthorship*, serán calificados como 0.1, lo mismo aplica al campo concatenado de *scientificName*, por lo que los datos para los campos de referencia del nombre científico como *taxonomicStatus*, *nomenclaturalStatus*, *acceptedNameUsage* o *nameAccordingTo*, en caso de que existan valores para estos campos en la tabla de datos de origen, se calificarán como inconsistentes, al no tener referencia con el catálogo.

Por otro lado, para aquellos registros con nombres científicos propuestos para ser evaluados por el proveedor de datos, se heredará la misma calificación a partir de la categoría taxonómica donde se hace la propuesta.

Por último, en el proceso de control de calidad, mientras se asignan y califican los valores procedentes de los catálogos taxonómicos, se recomienda utilizar campos de control, como *lastModifiedUser* para indicar quién realiza cambios en la "tabla de trabajo" y *lastModified* para señalar la fecha de la última modificación realizada. Estas asignaciones pueden llevarse a cabo de manera manual o automáticamente mediante funciones, dependiendo del software utilizado, lo que permite la asignación automática del nombre del usuario y la fecha en que se realizaron modificaciones en los valores de la "tabla de trabajo".

Perfil del personal

En términos generales, se requiere que el personal encargado de gestionar este dominio de datos en el control de calidad tenga conocimientos generales de taxonomía biológica y la capacidad de comprender y procesar información de este tipo. A continuación, se describen los rasgos más importantes que se deben poseer:

1. Conocimientos en taxonomía y nomenclatura
    a. Es importante que conozca los términos básicos de taxonomía biológica, incluyendo la estructura básica en la que deben escribirse los nombres científicos según los distintos códigos de nomenclatura, así como la forma en que se asocian las distintas categorías taxonómicas.
    b. Debe ser capaz de identificar las terminaciones establecidas en los códigos de nomenclatura para categorías de clasificación superior (phylum, clase, orden y familia) con el fin de asignarlas correctamente.
    c. Debe tener la capacidad de identificar el grupo biológico general al que pertenecen las bases de datos sobre las cuales se realizará el control de calidad, para así determinar el proceso de revisión con los catálogos elegidos por los proveedores de datos.
2. Conocimientos en bases de datos taxonómicas y catálogos
    a. Debe ser competente en la correcta identificación de las distintas categorías taxonómicas (reino, phylum, clase, orden, familia, género, subgénero, epíteto específico, rango taxonómico, epíteto infraespecífico, autor del nombre científico, estatus taxonómico, etc.) en diferentes catálogos para asignar correctamente la información en las bases de datos en las que se realiza control de calidad.
    b. Debe comprender cómo identificar la manera en que se debe atomizar o concatenar los campos que así lo quieran según las reglas de nomenclatura taxonómica.
3. Capacidad de análisis e integración

    a. Es importante identificar que, en taxonomía, todos los campos están asociados entre sí, y la asignación de información revisada y validada a cada uno debe tener coherencia interna entre los campos para un mismo nombre científico.
4. Revisión y validación minuciosa
    a. Dado que el nombre científico otorga una identidad al ejemplar, es necesario que el personal asignado a estas tareas sea capaz de verificar la información asignada para evitar errores de proceso que podrían generar combinaciones de nombres o categorías inexistentes.
5. Capacidad de investigación e iniciativa
    a. Es preferible que el personal tenga un interés genuino en investigar y revisar nuevos catálogos para complementar los ya existentes. Además, deben estar dispuestos a proponer nombres científicos para su validación al considerar las opciones disponibles por medio de la búsqueda de artículos científicos especializados en la descripción de nuevas especies o revisiones taxonómicas.
6. Aptitudes para el trabajo en equipo
    a. Debe ser capaz de fomentar un ambiente de trabajo colaborativo con los analistas líder y subalternos.
    b. Debe estar dispuesto a dialogar para enriquecer las metodologías establecidas en el control de calidad y proponer mejoras con justificaciones adecuadas.
    c. Debe mantener una actitud de confianza al comunicar errores en los catálogos, datos a revisar y metodologías de control de calidad, al mismo tiempo que propone soluciones para resolver estas situaciones.

Asignación de roles

Respecto a la asignación de roles en el control de calidad de los datos del dominio de taxonomía, están principalmente involucrados un analista líder y un analista de datos. Además, se requiere un coordinador y un analista de gestión de bases de datos. Las funciones principales se describen a continuación:

1. Coordinador.
2. Analista en metodología y estadística.
3. Analista líder de datos taxonómicos:
    a. Analiza en general la tabla de trabajo generada por el analista de metodología y estadística.
    b. Detecta problemáticas principales presentes en los datos a revisar.
    c. Establece estrategias y metodologías para el control de calidad
    d. Revisa los datos validados por el analista de datos de taxonomía.
4. Analista de datos (taxonomía)
    a. Encargado de la revisión de los datos específicos, en este caso taxonomía.

b. Realiza la comparación con catálogos para asignar la información validada para los registros y califica la consistencia de la información original en relación con la consulta en los catálogos correspondientes.
c. Detecta inconsistencias y propone soluciones que serán revisadas por el encargado de la colección.

Fechas

Dentro del dominio de datos de colecta se incluyen los valores fecha, que otorgan un contexto temporal al registro respecto al evento de colecta en sí. Este tipo de dato incluye tanto las fechas del evento de colecta como las fechas de identificación, y según la recomendación de CONABIO (2019), ambos eventos temporales pueden agruparse por ser el mismo tipo de dato. Por lo que su revisión y validación en el control de calidad es similar, ya que en la práctica del control de calidad realizado por la DGRU se ha observado que pueden aplicarse el mismo diseño de base de datos y uso de estándares para este tipo de datos.

La finalidad de llevar a cabo el control de calidad de este tipo de información recae en la necesidad de tener valores homogéneos y comparables entre sí, siguiendo las normas establecidas por el estándar Darwin Core. Esto permite que la información contenida pueda ser sujeta a búsquedas más precisas, reduciendo el número de variantes de una misma fecha y, a su vez, permite identificar fechas inconsistentes respecto a la forma en que están descritas (DGRU, 2022).

Estándares de datos

De manera similar a los datos taxonómicos, los datos procedentes de los registros de ejemplares biológicos están cubiertos por el estándar de Darwin Core (Darwin Core Task Group, 2009) (DGRU, 2019). En este estándar, los datos de la fecha del evento (en lo subsecuente denominado como fecha de colecta) se encuentran en la sección denominada "Evento", y en los datos de la fecha de identificación se ubican en un campo en la sección "Identificación".

En el caso de los datos de fecha, se emplea la Norma ISO 8601-1:2019 (International Organization for Standardization [ISO], 2019) como estándar para escribir el formato de las fechas y el tiempo. Este formato está enfocado en usar un patrón común para establecer los diferentes formatos de fechas provenientes de los registros de ejemplares, reduciendo así las variantes de una misma fecha de colecta o identificación. El formato básico se compone de año, mes y día (AAAA-MM-DD), con la posibilidad de incluir la hora en que se lleva a cabo el evento.

Para ilustrar, una fecha descrita con año, mes, día y hora de colecta, su formato debe ser: AAAA-MM-DDTHH:MM, donde AAAA representa los cuatro dígitos del año, MM indica

los dos dígitos del mes y DD los dos dígitos del día, y la T separa la fecha de la hora, que se indica con dos dígitos cada uno para la hora (HH) y los minutos (MM).

Algunas consideraciones importantes respecto al formato según la norma ISO al ingresar valores en el estándar Darwin Core (Darwin Core Maintenance Group, 2021) son:

- No deben incluirse notaciones para indicar la ausencia de algún elemento de la fecha o tiempo (hora) de colecta o identificación, simplemente el valor no se indica en los campos concatenados y se dejan vacíos los campos atomizados, según el caso.
- Para los campos "month" y "day" se establece la norma de que los números deben ser enteros. En el caso de valores que representen meses o días menores a 10, deben escribirse sin cero al inicio. Por ejemplo, el mes de enero se representa como "1", no como "01".
- El símbolo "–" (guion corto) se utiliza para separar los elementos año, mes y día (AAAA-MM-DD), mientras que el símbolo "/" (diagonal) se emplea para denotar los intervalos entre dos fechas. Por ejemplo, del 18 al 19 de julio de 2023 se indica como "2023-07-18/19".
- La letra "T" se usa para separar la hora en formato de 24 horas de la fecha del evento. Por ejemplo, "2023-07-19T15:00".
- Se debe mantener la relación entre días y meses. Por ejemplo, no puede existir una fecha descrita como 31 de abril, ya que abril solo cuenta con 30 días. Lo mismo se aplica a los años bisiestos, y es necesario verificar que las fechas que corresponden al 29 de febrero estén registradas en un año bisiesto.

Base de datos

El diseño de la base de datos sigue criterio basado en los campos correspondientes del estándar Darwin Core (Darwin Core Task Group, 2009) para este tipo de datos de colecta. En el caso de las fuentes de los datos, al igual que los datos taxonómicos, estos pueden tener distintos formatos, pero es indispensable mantener el formato original de las fechas de colecta o identificación al momento de importar la información a la base de datos para preservar la integridad de los datos.

En el Cuadro 5 se muestran los campos principales que provienen del estándar de Darwin Core (Darwin Core Maintenance Group, 2021) en las categorías "Evento" e "Identificación" utilizados en la creación de la base de datos. La mayoría de los campos son opcionales debido a que se prefiere la publicación del registro, incluso si no se cuenta con datos de fecha. Además, varios campos dependen de un campo principal, de manera similar a los datos de taxonomía, donde algunos campos representan concatenaciones y otros valores atomizados. Como resultado, algunos campos pueden quedar vacíos, por ejemplo, si una fecha de colecta no tiene más que el año y el mes, el campo "day" permanecerá vacío.

Cuadro 5. Campos de las clases "Evento" e "Identificación" del estándar Darwin Core empleados en el control de calidad de fechas de colecta e identificación

Tipo Campo	*Nombre campo Darwin Core*	*Estatus en control de calidad*
Campo de registro	eventDate	Opcional
	eventTime	Opcional
	startDayOfYear	Opcional
	endDayOfYear	Opcional
	year	Opcional
	month	Opcional
	day	Opcional
	verbatimEventDate	Opcional
	dateIdentified	Opcional
Campos de control	lastModified	Obligatorio
	lastModifiedUser	Obligatorio
Identificadores	occurrenceID	Obligatorio
	uuid	Obligatorio
	datasetID	Obligatorio

Nota. Para todos los "campos de registro" en la base de datos, se incluyen campos con el mismo nombre que el campo de registro, seguido de un sufijo que indica los valores originales (sufijo "_o"), las calificaciones (con sufijo "_qi") y los permisos o banderas (sufijo "_pub").

Fuente: Elaboración propia con base en la información de *Darwin Core Quick Reference Guide* (Grupo de Mantenimiento Darwin Core, 2021).

Herramientas de datos

En el subdominio de datos de colecta, se emplean principalmente consultas de comparación y actualización por medio de la identificación de patrones de formato en las fechas con la finalidad de homogeneizar de acuerdo a la norma ISO 8601-1:2019 (ISO, 2019). Durante el proceso de limpieza de datos, también se emplean consultas de atomización o concatenación.

Para llevar a cabo estos procesos, se puede emplear cualquier gestor de base de datos o programas similares que permitan realizar comparaciones internas dentro de los datos de la misma base. Asimismo, se puede crear una "tabla de trabajo" que contenga los datos para fechas y, por medio de ella, realizar su revisión.

En el control de calidad de los datos de fechas se tienen contempladas cinco fases principales: identificación de los campos fuente de los datos de fecha, evaluación del formato iso, limpieza de datos y estandarización al formato ISO, asignación de valores procesados y atomización de valores procesados. A continuación, se describen las actividades principales en cada una de estas fases.

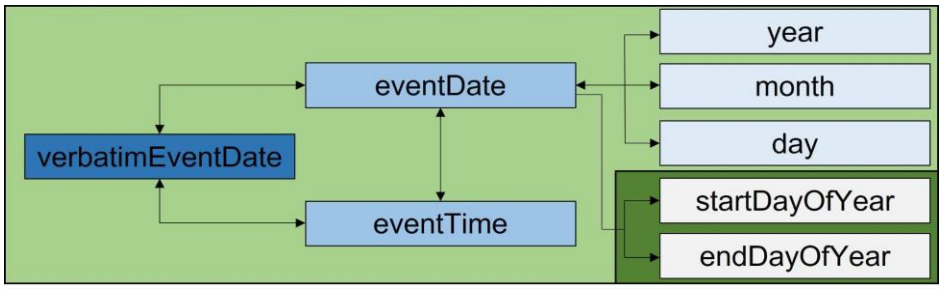
1. Identificación de los campos fuente de los datos de fecha

a. En el caso de los datos de fecha de identificación, se emplea el único campo "dateIdentified" disponible para el procesamiento y validación de la información. Esta fecha corresponde al último evento de determinación y está directamente relacionada con la última determinación taxonómica registrada.
b. Para los datos de fecha de colecta, existen distintas formas en las que puede presentarse la información en las bases de datos de origen, pudiendo estar disponible en uno o más campos. Las más comunes son:
    i. Fecha original (verbatimEventDate) y su transformación en formato ISO (eventDate), atomizada en los campos "year", "month" y "day".
    ii. Solo presente en el formato original (verbatimEventDate), sin atomización.
    iii. Solo en el formato ISO (eventDate) sin atomizar.
    iv. Solo en los campos atomizados, con o sin formato ISO (year, month y day).

    Es necesario que se comprenda la forma en la que están relacionados los campos que conforman la fecha de colecta dentro del estándar de Darwin Core. Los campos pueden integrarse en cuatro niveles de agrupación según el tipo de información que albergan de la misma fecha. Estos son:
        - *verbatimEventDate*
        - *eventDate* y *eventTime*
        - *year*, *month* y *day*
        - *startDayOfYear* y *endDayOfYear*

    Los primeros dos niveles contienen los datos de forma concatenada (*verbatimEventDate, eventDate* y *eventTime*), mientras que para el tercero se muestra de forma atomizada (*year*, *month* y *day*). Por lo que se pueden completar entre sí los datos presentes en los tres primeros niveles mediante atomización o concatenación de los datos en los campos. En contraste, para los campos del cuarto nivel (*startDayOfYear* y *endDayOfYear*), debido a la definición del campo en el estándar, pueden obtenerse a partir de los campos de los niveles 1 al 3. Sin embargo, no se puede obtener el dato completo de la fecha de colecta, al carecer estos campos de la referencia del año cuando se realizó la colecta. En la Figura 4 se ilustra un ejemplo de las relaciones entre los distintos campos que conforman la fecha de colecta y cómo se puede conformar un campo concatenado a partir de otros atomizados.

Figura 4. Relación de asociación entre los campos de la fecha del evento (colecta) en el estándar Darwin Core



Nota. El color del recuadro de cada campo representa el nivel de agrupación en función de los datos que albergan sobre la fecha de colecta. La intensidad del color disminuye a medida que desciende el nivel de agrupación. Las flechas indican el grado de relación entre los campos, ya sea de manera unilateral o bilateral. El color de fondo refleja cómo se relacionan los campos en la realización de concatenaciones o atomizaciones para completar datos.

2. Evaluación del formato ISO
    a. Una vez identificados los campos de origen que contienen la fecha, se debe asignar a los campos de valores originales (con el sufijo "_o"), siguiendo el estándar, para su posterior procesamiento. En el caso de la fecha de colecta, se utiliza el campo *verbatimEventDate*, mientras que para la fecha de identificación se emplea *dateIdentified*. El uso del campo con el sufijo "_o" es necesario para indicar que se trata del valor original que se someterá a revisión, ya que no existe un campo con el prefijo "*verbatim*" para *dateIdentified*.
    b. La evaluación del formato de la fecha original con respecto a lo establecido en la norma ISO 8601-1:2019 (ISO, 2019) se aplica a ambos tipos de fechas (colecta e identificación) en sus campos respectivos.

       Para corroborar si las fechas cumplen con el formato ISO, se analizan sus elementos por medio de consultas de búsqueda de patrones, utilizando expresiones regulares. De esta manera, se pueden identificar los elementos que las conforman las fechas, como el año, mes, día y hora, y asegurarse de que haya coherencia interna en el contexto del calendario que representan.

       El uso de las expresiones regulares permite identificar cuáles son las fechas que tienen más de un elemento, a través de la detección de las separaciones con guiones (-). Esto conduce a la identificación de tres niveles de precisión para la fecha representada, dependiendo del número de guiones presentes en la fecha:
        i. 2 guiones: la fecha incluye año, mes y día.
        ii. 1 guion: la fecha contiene año y mes.
        iii. Ningún guion: la fecha solo incluye el año.

       Además, se puede utilizar otra expresión regular que emplee en la búsqueda la letra "T" para separar el valor de la hora y el uso del signo de dos puntos (:) para distinguir la hora y los minutos.
    c. Una vez que se han identificado los elementos en formato ISO que componen la fecha, se procede a comparar la longitud de los valores individuales. Este

proceso se lleva a cabo mediante consultas de atomización que permiten obtener cada elemento de forma independiente. En este sentido, se aplican las siguientes consultas de expresiones regulares:

i. Se verifica que el primer elemento de la fecha sea de cuatro dígitos (año: AAAA).
ii. Los segundos y terceros elementos deben consistir en dos dígitos cada uno (mes: MM y día: DD).
iii. En caso de que hay información de tiempo, se requiere que se con dos dígitos tanto para la hora como para los minutos (horas: HH y minutos: MM).

d. A continuación, se identifican los rangos de valores que pueden poseer los elementos de la fecha.
    i. Año: el valor máximo (el más reciente) debe ser igual o menor al año en curso en el que se está realizando el control de calidad. Por ejemplo, si el control de calidad se realiza en el año 2023, las fechas con años mayores serían inconsistentes ya que no han ocurrido aún. Por otro lado, el valor mínimo (el más viejo) puede ser variable y, en general, dependerá del conjunto de datos en revisión y de las observaciones del proveedor de datos. Por ejemplo, en una colección con registros históricos, un año como 1885 podría ser considerado correcto, mientras que en una colección que tiene 20 años de existencia, ese mismo año sería inconsistente.
    ii. Mes: el valor debe ser entre 1 y 12, puesto que representan los doce meses del año.
    iii. Día: el valor puede variar entre 1 y 31. Sin embargo, el rango debe ser consistente con el mes de referencia de la fecha. Por ejemplo, enero (el primer mes) tiene 31 días, mientras que abril (el cuarto mes) tiene 30 días. Existe una consideración especial para el 29 de febrero, que solo es válido en años bisiestos, es decir, en aquellos años que incluyen un 29 de febrero en su calendario.
e. Se debe verificar si la fecha que se está revisando corresponde a un intervalo, es decir, si existe un día de inicio y otro de finalización del evento de colecta o de identificación. Para lograr esto, se puede utilizar una consulta con expresión regular que identifique el símbolo diagonal (/) dentro del valor de la fecha, y es crucial que la presencia de esta barra diagonal corresponda con el uso adecuado definido por la norma ISO 8601-1:2019 (ISO, 2019).

    Todos estos pasos se deben considerar para cada fecha que conforma el intervalo, y es necesario asegurarse de que ambas fechas existan y de que la primera sea menor (más antigua) que la segunda. Si el formato es correcto, se considera como consistente. Sin embargo, si no cumple con los requisitos, se puede realizar la limpieza de los datos (Fase 3) para identificar la razón de la inconsistencia y, en

su caso, estandarizar al formato ISO para considerarlo como una fecha consistente.

f. Por otro lado, se debe tener en cuenta que una fecha resulta inconsistente si solo incluye mes y día, pero no indica año de colecta o identificación. Sin embargo, una fecha que solo presenta año o año y mes es consistente, aunque carezca de dato para día. Del mismo modo, la presencia o ausencia del dato de hora en la fecha no la hace inconsistente, a menos que su valor sea incorrecto por sí mismo (por ejemplo, un valor que supere las 24 horas o que contenga más de 59 minutos).
g. Finalmente, tras revisar los aspectos mencionados, se puede establecer la fecha como consistente con el formato ISO y se puede asignar a los campos procesados correspondientes (Fase 4). Mientras que, si la fecha resulta inconsistente, es posible realizar una limpieza de datos y estandarizarla al formato ISO (Fase 3).

3. Limpieza de datos y estandarización al formato ISO
    a. Aunque una fecha pueda no cumplir con el formato ISO de origen, es posible que sea consistente, por lo que se puede llevar a cabo una limpieza de datos y realizar la posterior estandarización al formato ISO. Las razones principales por las que las fechas no corresponden de origen con el formato ISO incluye:
        i. Presencia de espacios en blanco en los elementos de la fecha.
        ii. Uso de distintos símbolos para delimitar los elementos de la fecha, como barras diagonales, preposiciones, espacios, etc.
        iii. Representación de elementos faltantes con notaciones no válidas, como NA, ND, @, 99, etc.
        iv. Diversos ordenamientos de los elementos de la fecha, como comenzar por el día o el mes.
        v. Abreviaturas en los años, como utilizar solo dos números para indicar el año que corresponde a una década en particular, como 99 en lugar de 1999.
        vi. Uso de palabras o números romanos para denominar elementos de la fecha, como utilizar el nombre del mes en lugar de un número.
        vii. El idioma en el que están escritas las fechas, por ejemplo “04/05/1999” en inglés podría corresponder al 5 de abril de 1999, mientras que en español corresponde al 4 de mayo de 1999.

    Para identificar este tipo de elementos, se puede hacer uso de expresiones regulares a partir de la exploración de los patrones en los datos de las fechas. Por ejemplo, si se identifica un mes escrito como “nov” (noviembre) se ocupan consultas de actualización para convertir el valor en 11 (el número correspondiente al mes).

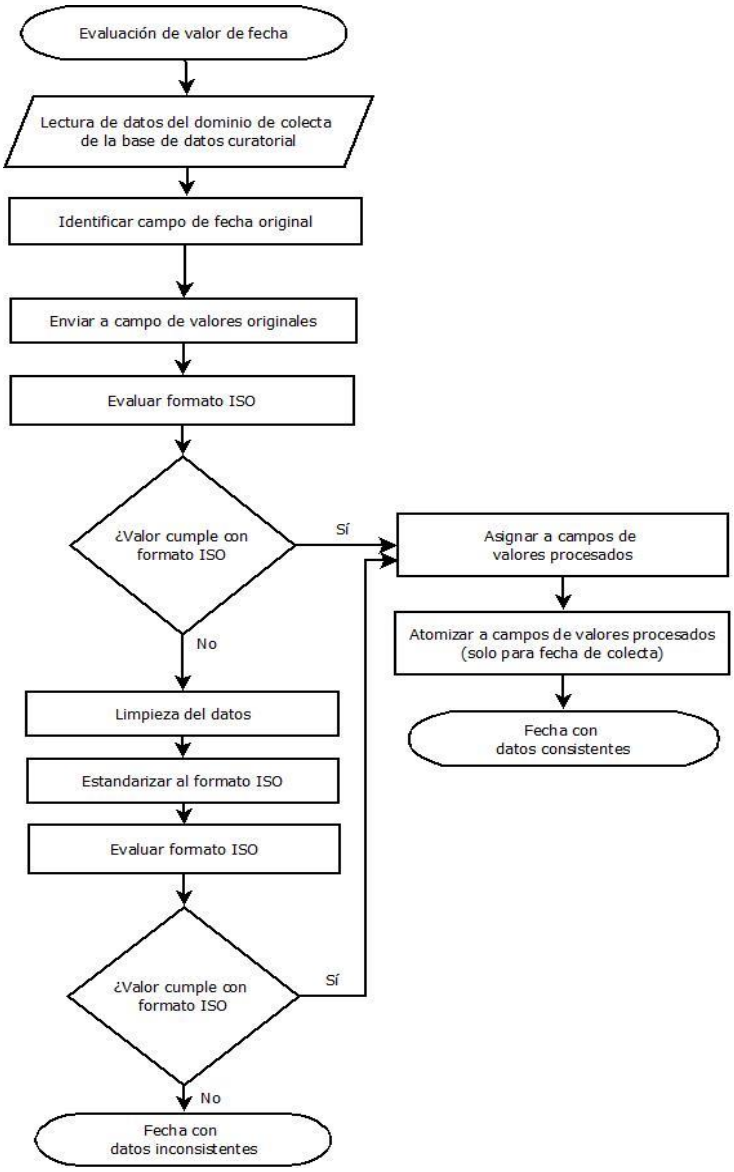
    Una vez que se han codificado correctamente los elementos de la fecha, se ordenan y estandarizan con el formato de la norma ISO (AAAA-MM-DDTHH:MM). Luego se revisa la consistencia lógica tomando en cuenta los

criterios de la Fase 2, por lo cual la fecha puede ser considerada como consistente, si no se cumple alguno de ellos se considera inconsistente.

4. Asignación de valores procesados
    a. Cuando se determina que la fecha cumple con el formato ISO y es consistente, se puede asignar el valor procesado a los campos correspondientes del estándar Darwin Core, usando las versiones sin sufijo de estos campos. En el caso de la fecha de colecta se emplea el campo *eventDate*, mientras que para la fecha de identificación se emplea *dateIdentified*.
    b. Por otro lado, únicamente para la fecha de colecta, si su nivel de precisión incluye el día de colecta, se deben asignar los valores correspondientes a los campos *startDayOfYear* y *endDayOfYear*, dichos valores corresponden al día del año en que comienza y termina el evento de colecta, respectivamente. En caso de que no exista un intervalo de fechas, solo se asignará el dato para *startDayOfYear* y el campo *endDayOfYear* quedará vacío. La asignación de estos datos se puede realizar por medio de distintas funciones o fórmulas, dependiendo del software utilizado, por ejemplo, la función EXTRACT (DOY) en el software PostgreSQL (PostgreSQL Tutorial Website, 2022).
5. Atomización de valores procesados
    a. Para las fechas de colecta, es necesario atomizar los datos para los campos *year*, *month*, *day* o *eventTime,* según sea el caso del nivel de precisión contenida en la fecha, es decir, que, si una fecha no tiene dato para día, el campo *day* queda vació. Para la asignación se pueden emplear las mismas consultas de atomización usando como referencia el símbolo guion (-) (para *year*, *month* y *day*) o la letra "T" (para *eventTime*) con el uso de expresiones regulares.
    b. Adicionalmente, se debe verificar que, para los campos *month* y *day,* los valores menores a 10 se ingresen únicamente con un dígito. Por ejemplo, la fecha "2023-07-05" debe tener los datos 7 y 5 en los campos *month* y *day,* respectivamente.
    c. En el caso de los eventos de colecta que corresponden a un intervalo de fechas, se sugiere atomizar los datos de la fecha inicial del rango para asignarlos a los campos *year*, *month*, *day* o *eventTime,* según corresponda.
    d. Por otro lado, para la fecha de identificación, no se realiza la atomización de la información contenida en la fecha, ya que no existen dichos campos en el estándar Darwin Core.

Las fases del proceso de control de calidad para los datos de fechas se resumen en el diagrama de flujo presente en la Figura 5.

Figura 5. Diagrama de flujo de la revisión de los datos del subdominio de fechas (dominio de datos de colecta)



Nota. Se muestran las fases principales de la identificación, estandarización con el formato ISO, limpieza de datos, asignación y atomización de datos. Los campos empleados como receptores de los valores originales de fechas son *verbatimEventDate* para colecta y *dateIdentified* para identificación. Los campos para asignar los valores procesados de fecha son: *eventDate* para colecta y *dateIdentified* para identificación. La atomización de datos (*year*, *month*, *day* y *eventTime*) que conforman la fecha solo aplica para fechas de colecta (*eventDate*).

Evaluación de la calidad de datos

En el proceso de evaluación de la calidad de los datos, se realiza una asignación de calificaciones a los campos de fechas de colecta e identificación. Esto se logra comparando los valores de los campos originales (identificados con el sufijo "_o") con los valores obtenidos después de la revisión y la asignación del formato de la norma ISO en una "tabla

de trabajo" (campos sin sufijo). Para determinar estas calificaciones, se emplean los campos con el sufijo "_qi", y se asigna una de las calificaciones que se describen en la sección 3 de este manual.

Cada campo de este subdominio de datos tiene su propia calificación. El cuadro 6 muestra las calificaciones que son aplicables para todos los campos con datos de fechas. Los dos grupos principales de calificaciones incluyen consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Cuadro 6. Lista de calificaciones (qi) empleadas en el control de calidad de datos de fechas

	Calificación (qi)	*Situación*	*Ejemplos*
consistente	1	Consistente de origen: el campo original y procesado son iguales	eventDate_o: 2022-09-11 eventDate: 2022-09-11
	0.9	Modificado por variantes de escritura (por ejemplo, digitación, ortografía, cambio de minúsculas o mayúsculas, completar y espacios en blanco). La identidad del valor se mantiene.	eventDate_o: 1978**/**02**/**28 eventDate: 1978**-**02**-**28
	0.8	Dato asignado durante el proceso de control de calidad, es decir, que no está presente en el campo original.	Para la fecha 2022-01-20 el valor de día debe ser 20. startDayOfYear_o: sin dato startDayOfYear: **20**
	0.7	Modificado por estandarización respecto a la norma ISO.	Para la fecha 14 de junio de 1925 month_o: **junio** month: **6**
inconsistente	0.4	Requiere la resolución del proveedor de datos tras realizar la revisión en el control de calidad.	a) Año abreviado: 14-01-**89** Dependiendo del contexto podría ser: **1989**-01-14 o **1889**-01-14 b) Cuando la fecha de colecta es más reciente que la identificación
	0.1	No fue posible encontrar relación lógica ni con la definición del campo, según el estándar, ni se puede corroborar con otras referencias, por lo que resulta inconsistente.	Fechas que no existen: 2018-**02-29** (el año 2018 no fue bisiesto) o incompletas **06-10** (falta un año)
nulo	0	El dato original se eliminó del campo.	eventDate_o: Helia Bravo eventDate: sin dato
	Vacío (sin dato)	Cuando no existe dato para revisar.	Cuando una fecha no tiene valor para el día el campo day permanece vacío day_o: sin dato day: sin dato

Nota. Se muestra la codificación para cada valor de calificación (qi), se describe la situación correspondiente y se proporcionan ejemplos de los casos en los que se aplica dicha calificación. Las calificaciones se dividen en categorías: consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Fuente: Elaboración propia a partir de los nombres de los campos de Darwin Core (Grupo de Mantenimiento Darwin Core, 2021).

Sin embargo, después de calificar individualmente cada campo, es fundamental comparar los valores en su conjunto, dada la relación que existe entre los campos que conforman de

fecha, principalmente en el caso de la fecha de colecta. Por ejemplo, existen algunos casos en los que la fecha completa (*verbatimEventDate* o *eventDate*) puede ser inconsistente, pero no todos sus elementos lo son, por lo que es necesario indicarlo correctamente con las calificaciones de cada campo para identificar qué elementos son los que hacen inconsistente a la fecha. Algunos ejemplos de dicha situación se muestran en el Cuadro 7.

Cuadro 7. Ejemplos de calificación para los campos de fecha de colecta

Caso	*Verbatim Event Date*	*Event Date*	*Event Time*	*year*	*month*	*day*	*Explicación*
Año incompleto, sin referencia específica	05-06-200 **0.1**	200-06-05 **0.1**	vacío **sin dato**	200 **0.1**	6 **1**	5 **1**	*year* es inconsistente, ya que no es posible especificar a qué año de la década de los 2000 corresponde.
Año no bisiesto	29 de febrero 2018 a las 3 pm **0.1**	2018-02-29T15:00 **0.1**	15:00 **1**	2018 **1**	2 **1**	29 **0.1**	El año 2018 no fue bisiesto.
Fecha inexistente	31/04/1993 16 hrs **0.1**	1993-04-31T16:00 **0.1**	16:00 **1**	1993 **1**	4 **1**	31 **0.1**	El día 31 de abril no existe en el calendario.

Nota. Se muestran algunos casos de fechas inconsistentes y cómo calificar los campos que contienen las inconsistencias. Las columnas hacen referencia a los campos que integran la fecha, el dato que ejemplifica y la calificación (resaltadas en negritas) que les corresponde según el caso. En la columna "Explicación" se indica si el valor de la fecha en su conjunto es consistente o no, y la razón de ello. No se incluyen los campos *startDayOfYear* ni *endDayOfYear*, ya que al ser fechas inconsistentes estos valores no se pueden asignar con certeza.

Fuente: Elaboración propia a partir de los nombres de los campos de Darwin Core (Grupo de Mantenimiento Darwin Core, 2021).

Además, para el campo de fecha de identificación (*dateIdentified*), al no existir campos atomizados en el estándar Darwin Core, su calificación se mantiene de manera individual, siendo inconsistente solo si la fecha no existe o es errónea en su codificación original y no puede ser estandarizada con la norma ISO.

Una vez evaluadas ambas fechas (colecta e identificación) y en caso de que ambas estén presentes en el mismo registro, se realiza la comparación entre ambos valores para determinar si se cumple la siguiente relación: *eventDate* ≤ *dateIdentified*. Dicha relación parte de la premisa de que un evento de colecta debe ocurrir antes o al mismo tiempo que un evento de identificación, ya que resultaría incoherente que se identifique la especie de un ejemplar antes de ser colectado.

De modo que, si se presenta el caso en el que la fecha en el campo *dateIdentified* es anterior a la fecha en el campo *eventDate*, ambas fechas deben calificarse como 0.4 (inconsistentes). Esto se hace con la finalidad de que sea revisada por el proveedor de datos, ya que es posible

que las fechas estén invertidas o alguna de ellas presenten un error de origen o captura en la información.

Por último, en el proceso de control de calidad, mientras se asignan y califican los valores procedentes de los catálogos taxonómicos, se recomienda utilizar campos de control, como *lastModifiedUser* para indicar quién realiza cambios en la "tabla de trabajo" y *lastModified* para señalar la fecha de la última modificación realizada. Estas asignaciones pueden llevarse a cabo de manera manual o automáticamente mediante funciones, dependiendo del software utilizado, lo que permite la asignación automática del nombre del usuario y la fecha en que se realizaron modificaciones en los valores de la "tabla de trabajo".

Perfil del personal

En líneas generales, se requiere que el personal encargado de este subdominio de datos en el control de calidad tenga conocimientos generales de biología y pueda entender y procesar la información del evento de colecta. A continuación, se describen los rasgos más importantes.

1. Conocimientos en colecciones biológicas y eventos de colecta
    a. Es importante que conozca los elementos básicos de los eventos de colecta (fechas de colecta) e identificación (fecha de identificación) para poder identificarlos claramente dentro los campos correspondientes.
    b. Reconocer los distintos formatos de fechas existentes dentro de las bases de datos.
2. Conocimientos sobre las normas para establecer el formato de fecha
    a. Estar familiarizado con la implementación del uso de la norma ISO 8601-1:2019 (ISO, 2019).
    b. Identificar la manera en que se debe atomizar o concatenar los campos que así lo quieran según las especificaciones del estándar de datos Darwin Core (Darwin Core Maintenance Group, 2021).
3. Capacidad de análisis e integración
    a. Es importante identificar que los elementos que conforman las fechas están interrelaciones entre sí, y la asignación de información revisada y validada a cada uno debe mantener coherencia entre los campos atomizados (por ejemplo, "year", "month" y "day") y los campos concatenados (como "eventDate").
    b. Identificar patrones en los elementos de la fecha y sus separadores para la aplicación de expresiones regulares.
4. Revisión y validación minuciosa
    a. Dado que los formatos de fechas pueden variar debido a la codificación del software utilizado, se requiere atención al momento de realizar las integraciones de los valores originales en las "tablas de trabajo", así como su correcta codificación mientras se realiza en control de calidad.

5. Capacidad de investigación e iniciativa
    a. Se valora la disposición para investigar nuevas formas de revisar la información, como nuevas consultas, herramientas y programas para analizar de forma más eficiente la información, así como de documentar dichos hallazgos.
6. Aptitudes para el trabajo en equipo
    a. Debe ser capaz de fomentar un ambiente de trabajo colaborativo con los analistas líder y subalternos.
    b. Debe estar dispuesto a dialogar para enriquecer las metodologías establecidas en el control de calidad y proponer mejoras con justificaciones adecuadas.
    c. Debe mantener una actitud de confianza al comunicar errores en los catálogos, datos a revisar y metodologías de control de calidad, al mismo tiempo que propone soluciones para resolver estas situaciones.

Asignación de roles

Respecto a la asignación de roles en el control de calidad de los datos del subdominio de fechas (colecta e identificación), se involucra principalmente a un analista líder de datos y un analista de datos especializado en fechas. De igual forma se necesita de un coordinador y un analista de gestión de bases de datos. Las funciones principales se describen a continuación:

1. Coordinador.
2. Analista en metodología y estadística.
3. Analista líder de datos de colecta:
    a. Análisis general de la tabla de trabajo generada por el analista de metodología y estadística.
    b. Detección de problemáticas principales presentes en los datos a revisar.
    c. Establecimiento de estrategias y metodologías a seguir en el control de calidad.
    d. Revisión de los datos validados por el analista de datos de colecta.
4. Analista de datos (fechas de colecta e identificación)
    a. Encargado de la revisión de los datos específicos, en este caso para fechas.
    b. Realiza la revisión de los valores originales de las fechas de colecta e identificación del formato y limpieza de datos para el control de calidad.
    c. Asigna la información validada para los registros y, al mismo tiempo, califica la consistencia de la información original con respecto a la revisión y estandarización conforme a la norma ISO 8601-1:2019 (ISO, 2019).
    d. Detecta inconsistencias y propone soluciones para que sean revisadas por el proveedor de datos.

Nombre de personas

Dentro del dominio de datos de colecta, se incluyen una serie de datos relacionados con nombres de personas, ya sea que se refieran a los colectores o a los identificadores de ejemplares, ya que las metodologías que se emplean para revisar estos datos son similares (Castillo *et al*., 2014). Además, en la práctica del desarrollo del control de calidad en la DGRU, se ha observado que revisar estos datos en conjunto permite corroborar la escritura de algunos nombres. Esto es común cuando solo existe un evento de identificación, ya que la persona que realiza esta identificación suele ser la misma que colectó el ejemplar.

El control de calidad aplicado a este tipo de valores está enfocado en reducir la cantidad de variantes de escritura de los nombres de los colectores o identificadores. Esto se logra mediante la detección de nombres incompletos, abreviaturas, espacios innecesarios, o la separación de distintos nombres en un equipo de colecta o identificación, entre otros aspectos. El objetivo es homogeneizar estos nombres para facilitar búsquedas más precisas, como la agrupación de registros pertenecientes al mismo colector o grupo de colecta dentro de la base de datos.

Estándares de datos

Los nombres de personas se encuentran definidos en el estándar de Darwin Core (Darwin Core Task Group, 2009) (DGRU, 2019), principalmente en dos secciones: "Identificación", que incluye los nombres que hacen referencia al identificador taxonómico del ejemplar en el campo *identifiedBy*, y la sección "Ocurrencia", que describe el campo *recordedBy*, refiriendo al nombre del colector. Para ambos campos, es posible incluir uno o varios nombres de personas, así como nombres de instituciones o grupos que llevan a cabo los eventos de colecta o identificación.

Dada la inherente heterogeneidad de los nombres de personas, el estándar considera como aspectos importantes a tener en cuenta el orden en el que aparecen de los nombres, por ejemplo, si empiezan por el nombre (o nombres) o por el apellido, y cómo realizar la separación entre los nombres de personas cuando se hace referencia a más de uno.

Con respecto a la forma de separar los nombres cuando hay más de un integrante, el estándar Darwin Core sugiere usar una barra vertical y espacio (|) entre cada uno (Darwin Core Maintenance Group, 2021). Sin embargo, también se pueden considerar otras maneras de realizar la separación de cada nombre por medio de signos de puntuación, como el punto y coma seguido de un espacio (;). En este aspecto, es importante utilizar el mismo criterio en la separación de los nombres para homogeneizar, en la medida de lo posible, los equipos de colecta o identificación. Sin embargo, esta acción no implica cambiar el orden de aparición de los nombres de personas en el equipo de colecta o identificación.

Base de datos

La creación de la base de datos se lleva a cabo utilizando los campos correspondientes del estándar Darwin Core (Darwin Core Task Group, 2009) para los campos de los colectores e identificadores. Por lo que es importante asegurarse de que en las bases de datos de origen se precise adecuadamente el campo fuente de esta información, como la identificación correcta de los denominadores de campo. Dado que el tipo de datos que corresponde a la definición del campo es muy similar, se debe evitar generar errores durante la importación de datos.

En el Cuadro 8 se muestran los campos provenientes del estándar de Darwin Core (Darwin Core Maintenance Group, 2021) en las categorías "Ocurrencia" e "Identificación", los cuales se utilizan en la creación de la base de datos. Tanto los campos *recordedBy* como *identifiedBy* son de carácter opcional, ya que dicha información puede estar o no disponible en el registro del ejemplar.

Cuadro 8. Campos de las clases "Ocurrencia" e "Identificación" del estándar Darwin Core empleados en el control de calidad de nombres de colectores e identificadores

Tipo Campo	*Nombre campo Darwin Core*	*Estatus en control de calidad*
Campo de registro	recordedBy	Opcional
	identifiedBy	Opcional
Campos de control	lastModified	Obligatorio
	lastModifiedUser	Obligatorio
Identificadores	occurrenceID	Obligatorio
	uuid	Obligatorio
	datasetID	Obligatorio

Nota: En todos los "campos de registro" de la base de datos se incluyen campos con el mismo nombre que el campo de registro, más un sufijo para indicar los valores originales (sufijo "_o"), las calificaciones (el sufijo "_qi") y sus permisos o banderas (con el sufijo "_pub").

Fuente: Elaboración propia con base en *Darwin Core Quick Reference Guide* (Grupo de Mantenimiento Darwin Core, 2021).

Catálogos

En el estándar Darwin Core, se establece que los campos *recordedBy* e *identifiedBy* pueden hacer referencia tanto a individuos como a grupos, organizaciones o instituciones (Darwin Core Maintenance Group, 2021).

Dado este este enfoque diversificado, resulta inviable contar con un único catálogo general que incluya los nombres de personas, puesto que, a diferencia de los catálogos taxonómicos, los posibles catálogos estarían especializados en los grupos particulares de organismos biológicos que estudian los colectores o identificadores. Por ejemplo, existen catálogos con nombres de colectores o identificadores del mundo, pero enfocados a colectas botánicas, como los catálogos de autores disponibles en el sitio web de Tropicos (Tropicos, 2023b), el Index of Botanists of the Harvard University Herbarium (The President and Fellows of

Harvard College, 2013), o el catálogo Global Plants de JSTOR (ITHAKA, 2000-2023). Estos catálogos son bases de datos de nombres de personas que fungen como colectores de ejemplares y, en algunos casos, también han realizado identificaciones taxonómicas.

Otras fuentes de información, como listados o artículos, pueden fungir como catálogos, los cuales corresponden a revisiones específicas realizadas con enfoques particulares de grupos biológicos, regiones geográficas o períodos temporales. Ejemplos de esto incluyen el listado de colectores de algas de México desde 1787 hasta 1954 (Godínez, 2008), la lista de autores e instituciones presentes en las colecciones mastozoológicas de México (Lorenzo *et al.*, 2006), el listado histórico de los principales colectores botánicos en México desde 1700 hasta 1930 (Rzedowski *et al.*, 2009), o el listado de colectores de plantas vasculares de México (Murguía-Romero *et al.*, 2022).

Sin embargo, dicha información podría no cubrir la totalidad de los nombres asociados a las colectas (ejemplares), debido a que, en algunos casos, los colectores o identificadores, según la experiencia observada en el control de calidad dentro de la DGRU, podrían ser estudiantes, pasantes, recién graduados, colaboradores en los viajes de campo o residentes locales, e incluso personal interno de instituciones que no figuren en bases de datos públicas. Por lo tanto, en estos casos, podría no existir una fuente de referencia confiable.

Además, cabe la posibilidad de que los proveedores de datos proporcionen listados de los colectores e identificadores específicos para eventos o colecciones particulares. La elección de una o más fuentes de referencia para los nombres de personas dependerá del grupo biológico bajo revisión y de la procedencia de la colección, y se acordará en colaboración con el proveedor de datos.

Herramientas de datos

En el contexto de la gestión de datos en este subdominio de colecta de datos se emplean consultas de comparación, de atomización, concatenación y actualización. El objetivo principal de estas acciones es garantizar la integridad de la información relativa a los campos *recordedBy* e *identifiedBy* en una "tabla de trabajo", que se construye a partir de los datos de origen. Esto resulta fundamental para mantener la integridad de los nombres de las personas y el orden original, especialmente cuando se trata de equipos de colecta o identificación. Por ejemplo, el primer nombre en la lista de colectores de un equipo de colecta debe mantenerse siempre en esa posición, aunque el valor del nombre sea homogeneizado.

Para llevar a cabo el control de calidad de los datos de nombres de personas se contemplan cinco fases principales de revisión, comparación, asignación y validación de la información. A continuación, se describen y mencionan las actividades principales en cada una de estas fases.

1. Identificación de campos fuente de datos de nombres de personas

a. Se deben identificar correctamente los nombres de las personas que corresponden a colectores y asignarlos al campo *recordedBy*, mientras que los nombres que correspondan a identificadores deben incluirse en el campo *identifiedBy*.
b. Dicha identificación es importante ya que es posible que los nombres de personas se repitan en ambos campos. Para evitar esta ambigüedad, se recomienda revisar el denominador de campo de la base de datos de origen antes de realizar la revisión de los datos originales, especialmente cuando difiere en sentido del estándar Darwin Core.

2. Establecimiento del formato y separación de nombres
    a. Como se mencionó anteriormente, el establecimiento de estos criterios debe ser abordado en conjunto con el proveedor de datos y debe tomar en cuenta las recomendaciones del estándar Darwin Core (Darwin Core Maintenance Group, 2021).
    b. Este formato debe mantenerse constante a lo largo de todas las revisiones de las entregas de datos nuevos de cada colección. Sin embargo, el formato puede variar entre proveedores de datos de entidades distintas.
3. Identificación de la forma de separación de los nombres de origen
    a. Existen diversas formas en las que se separan los nombres, como el uso de espacios simples (esto hace ambigua la delimitación de los nombres), o la utilización de signos de puntuación (como coma, punto y coma, guion, entre otros), conjunciones (como "y" o "e") e incluso otras palabras que delimitan al colector o identificador principal en relación con los secundarios o equipos de apoyo (como "con", "con el apoyo de").
4. Revisión de nombres de personas y criterios de homogeneización
    a. Los nombres de ambos campos (*recordedBy* e *identifiedBy*) se pueden atomizar mediante consultas con expresiones regulares que se basan en los elementos que separan los nombres (por ejemplo, espacios, coma, punto y coma, conjunciones etc.), con el fin de poder crear una "tabla de trabajo adicional" en la cual se puedan revisar de forma individual, principalmente cuando son varios nombres en los campos de *recordedBy* o *identifiedBy*. Es importante mantener la relación con el campo original y el orden que tiene dentro de la lista de nombres para un registro en particular (para los registros que tienen más de un colector o identificador).
    b. La revisión incluye la limpieza de datos, como la eliminación de espacios en blanco, la corrección de puntuación en abreviaturas, la conversión de mayúsculas y minúsculas, y la homogeneización propia de los nombres de personas. Esto puede lograrse utilizando fuentes especializadas para algunos nombres de colectores o identificadores (como se menciona en la sección "Catálogos"), o bien por medio del uso de herramientas especializadas en limpieza de datos, como OpenRefine (Delpeuch, 29 de diciembre de 2022)

(Muñoz *et al.*, 2016), que permiten realizar estas acciones de forma semiautomática y distinguir los patrones dentro de los nombres de personas.

c. En caso de no contar con catálogos o listas de nombres de personas, se pueden aplicar criterios desarrollados en la DGRU, esto puede incluir el uso de la misma base de datos como un autocatálogo de nombres, utilizando consultas de selección con parámetros o con la ayuda del programa OpenRefine (Delpeuch, 29 de diciembre de 2022) (Muñoz *et al.*, 2016) a fin de identificar posibles similitudes entre las variantes de un mismo nombre de persona. Una vez identificados los patrones de similitud, se pueden establecer criterios para homogeneizarlos, como se detalla a continuación:
    i. Si existen variantes de un mismo nombre de persona y se confirma que se refieren a la misma persona (por ejemplo, mediante la coincidencia de fechas o lugares de colecta), se puede utilizar la versión más común en la base de datos, seguida de la versión más específica (con nombres propios completos en lugar de iniciales), para completar y homogeneizar el nombre en cuestión.
    ii. Si se identifican claramente las iniciales de los nombres propios cuando están abreviados, estas iniciales pueden completarse con el nombre completo presente en la base de datos para dicho colector o identificador.
    iii. Es importante identificar la correcta escritura y acentuación de los nombres o apellidos para evitar errores en el procesamiento de los datos. Por ejemplo, "Souza" y "Sousa" son apellidos diferentes y no deben considerarse errores ortográficos.

5. Asignación y concatenación de los nombres de personas
    a. Una vez completada la limpieza y homogeneización de los nombres de personas, se deben concatenar los nombres pertenecientes a un mismo registro, siguiendo el orden original del campo correspondiente. Esto se realiza en caso de que se hayan separado en una "tabla de trabajo adicional" para revisar los nombres de forma individual. Si no se realizó este paso intermedio, la información homogeneizada se puede emplear directamente para asignarla a los campos procesados correspondientes. Para ambos casos, debe asignarse a los campos *recordedBy* e *identifiedBy* la información homogeneizada.
    b. La asignación de dichos valores homogeneizados se puede realizar por medio de las consultas de actualización dentro de la misma "tabla de trabajo" o desde la "tabla de trabajo adicional" a la "tabla de trabajo" principal.

Evaluación de la calidad de datos

La evaluación de los campos que contienen nombres de personas se realiza por medio de la asignación de calificaciones. Esta asignación se hace comparando los valores de los campos originales (con sufijo "_o") con los valores obtenidos de la revisión y homogeneización de

los nombres. La revisión puede involucrar la comparación con catálogos o listados especializados, así como la revisión interna de la misma base de datos, todo ello dentro de la "tabla de trabajo" que contiene los campos sin sufijo.

Para la asignación de las calificaciones, se emplean los campos con sufijo "_qi", en los c se asigna una de las calificaciones que se mencionan en la sección 3 de este manual.

En el caso de los campos de este subdominio de datos, se califica únicamente el valor final que se incluye en los campos *recordedBy* o *identifiedBy*, es decir, no existe una calificación por cada elemento que componga la información de dichos campos, como, por ejemplo, cuando hay más de un colector o identificador, sino que se califica toda la secuencia de los nombres en su conjunto. En el Cuadro 9 se muestran las calificaciones que son aplicables para todos los campos con información de colectores o identificadores. Los dos grupos principales de calificaciones se dividen en "consistentes" (mayores o igual a 0.6), "inconsistentes" (menores a 0.6) y "nulo" (sin dato).

Cuadro 9. Lista de calificaciones (qi) empleadas en el control de calidad de datos de nombres de personas

	Calificación (qi)	*Situación*	*Ejemplos*
consistente	1	Consistente de origen: el campo original y procesado son iguales.	recordedBy_o: Jerzy Rzedowski Rotter recordedBy: Jerzy Rzedowski Rotter
	0.9	Modificado por variantes de escritura (por ejemplo, digitación, ortografía, cambio de minúsculas o mayúsculas, completar y espacios en blanco). La identidad del valor se mantiene.	recordedBy_o: Mario Sousa S**a**nche**s** recordedBy: Mario Sousa S**á**nche**z**
	0.8	Dato asignado durante el proceso de control de calidad, es decir, que no está presente en el campo original.	El valor de colector estaba asignado en el campo "identifiedBy", en lugar de "recordedBy" y viceversa: recordedBy_o: Léia Scheinvar recordedBy: **Helia Bravo Holis** identifiedBy_o: Helia Bravo Holis identifiedBy: **Léia Scheinvar**
	0.7	Modificado por estandarización con respecto a un catálogo o listado de colectores o revisión interna de los nombres de la base de datos (cambia la identidad del valor).	Completar las iniciales de un nombre: recordedBy _o: H. Bravo H. recordedBy: Helia Bravo Holis
inconsistente	0.4	Requiere la resolución del proveedor de datos tras realizar la revisión en el control de calidad.	El colector Bravo Hollis puede referirse a dos personas diferentes (Margarita o Helia) por lo que se debe especificar.
	0.1	No fue posible encontrar relación lógica con la definición del campo según el estándar, ni se puede corroborar con otras referencias, lo que genera una inconsistencia.	Si existe algún nombre incompleto que no pueda ser corroborado, el valor es ilegible en la etiqueta de colecta o el valor escrito no corresponde con la definición del campo.
nulo	0	El valor original se eliminó del campo.	recordedBy_o: **1999-05-23** recordedBy: sin dato
	Vacío (sin dato)	Cuando no existe dato para revisar.	Si no existe valor de origen para uno de los dos campos con nombres de personas o no

	Calificación (qi)	*Situación*	*Ejemplos*
			se puede asignar uno, el campo correspondiente permanece vacío: identifiedBy _o: sin dato identifiedBy: sin dato

Nota. Se muestra la codificación para cada valor de calificación (qi), la situación que describe y algunos ejemplos de los casos en los que se ocupa dicha calificación. Las calificaciones se dividen en tres grupos: consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Fuente: Elaboración propia a partir de los nombres de los campos de Darwin Core (Grupo de Mantenimiento Darwin Core, 2021).

Como se mencionó anteriormente, la asignación de calificaciones para los campos *recordedBy* e *identifiedBy*, al realizarse directamente entre el valor del campo original (con sufijo "_o"), y el procesado (sin sufijo), no requiere alguna otra consideración adicional. Asimismo, la ausencia o presencia de información en algunos de los dos campos está sujeta a la base de datos original.

En el proceso de control de calidad, mientras se asignan y califican los valores procedentes de los catálogos taxonómicos, se recomienda utilizar campos de control, como *lastModifiedUser* para indicar quién realiza cambios en la "tabla de trabajo" y *lastModified* para señalar la fecha de la última modificación realizada. Estas asignaciones pueden llevarse a cabo de manera manual o automáticamente, dependiendo del software utilizado, lo que permite la asignación automática del nombre del usuario y la fecha en que se realizaron modificaciones en los valores de la "tabla de trabajo".

Perfil del personal

En general, se requiere que el personal encargado de este subdominio de datos en el control de calidad tenga conocimientos generales de biología y pueda ser capaz de comprender y procesar la información relacionada con el evento de colecta e identificación. A continuación, se describen los rasgos más importantes.

1. Conocimientos en colecciones biológicas y sus eventos de colecta e identificación
    a. Es importante que conozca los elementos básicos de los eventos de colecta (nombre del colector) e identificación (nombres de identificadores taxonómicos) para poder ubicarlos claramente dentro los campos correspondientes.
    b. Debe reconocer los distintos formatos de separación de los nombres para identificar correctamente el número de colectores o identificadores presentes en el registro.
2. Conocimientos sobre catálogos o listados con nombres de colectores o determinadores
    a. Debe estar familiarizado con las diferentes fuentes de nombres de personas considerados como colectores o identificadores de registros biológicos.

b. Debe identificar la manera en que se debe separar los nombres de personas, en caso de haber más de un colector o identificador, según las especificaciones del estándar de datos Darwin Core (Darwin Core Maintenance Group, 2021).

3. Capacidad de análisis e integración
    a. Es importante que pueda identificar que los nombres pueden variar dentro de una misma colección de registros y que puede haber sesgos generados por inconsistencias de origen en la escritura de los nombres de los colectores o identificadores.
    b. Tiene que poder identificar las posibles variantes de un mismo nombre de persona, especialmente cuando hay indicios de ser el mismo, como iniciales de nombres propios con apellidos y sus equivalentes de nombres completos, ya sea en la misma base de datos o en catálogos o listados de colectores o identificadores.
    c. Debe ser competente para la identificación de patrones de nombres de personas y las formas en las que se separan para la aplicación de expresiones regulares.
4. Revisión y validación minuciosa
    a. Tiene que ser capaz de revisar detalladamente el orden asignado a los nombres de personas dentro de los campos de *recordedBy* e *identifiedBy* y mantenerlo durante la revisión.
    b. Debe mantener el formato establecido para la separación de los nombres de personas en los campos de *recordedBy* e *identifiedBy*.
    c. Debe poder distinguir entre variantes de escritura de nombres de personas válidos y errores ortográficos, con apoyo de los catálogos o listados disponibles.
    d. Debe emplear lo más ampliamente posible las consultas y herramientas para realizar la comparación y revisión de los nombres de personas.
5. Capacidad de investigación e iniciativa
    a. Se valora la inclinación por la investigación y la búsqueda de nuevas fuentes de referencia para nombres de personas según los requerimientos de los registros a revisar, así como la de documentar dichos hallazgos.
6. Aptitudes para el trabajo en equipo
    a. Debe ser capaz de fomentar un ambiente de trabajo colaborativo con los analistas líder y subalternos.
    b. Debe estar dispuesto a dialogar para enriquecer las metodologías establecidas en el control de calidad y proponer mejoras con justificaciones adecuadas.
    c. Debe mantener una actitud de confianza al comunicar errores en los catálogos, datos a revisar y metodologías de control de calidad, al mismo tiempo que propone soluciones para resolver estas situaciones.

Asignación de roles

La asignación de roles en el control de calidad de los datos del subdominio de nombres de personas (colectores e identificadores) implica la participación fundamental de un analista líder y un analista de datos, específicamente enfocados en la información de nombres de personas. Además, se requiere la supervisión de un coordinador y de un analista de metodología y estadística. Las funciones principales se describen a continuación:

1. Coordinador.
2. Analista de metodología y estadística.
3. Analista líder de datos de colecta:
    a. Realiza un análisis general de la tabla de trabajo generada por el analista de metodología y estadística.
    b. Detecta problemáticas principales presentes en los datos a revisar.
    c. Establece estrategias y metodologías para el control de calidad.
    d. Revisa los datos validados por el analista de datos.
4. Analista de datos (nombres de personas: colectores e identificadores):
    a. Encargado de la revisión de los datos originales de los nombres de personas: colectores e identificadores.
    b. Realiza la identificación de nombres individuales cuando existen equipos de colecta o identificación.
    c. Asigna la información validada para los registros y, al mismo tiempo, califica la consistencia de la información original respecto a la revisión y consulta interna de la base de datos o con el uso de catálogos o listados de nombres de colectores o identificadores.
    d. Detecta inconsistencias y realiza propuestas para que sean revisadas por el proveedor de datos.

Geografía

La presencia de una especie se puede definir como una evidencia (observación o recolección) de un organismo en un espacio geográfico dado y en un momento determinado (Veiga *et al.*, 2014). Si se quiere analizar la información generada a partir de estos eventos de recolección, es esencial que los datos que indican la presencia de la especie sean de buena calidad. Veiga y colaboradores (2014) advierten que el uso de datos, sin considerar posibles errores, puede llevar a resultados imprecisos, información engañosa y, en consecuencia, a la toma de decisiones equivocadas.

La información geográfica es importante en diversas áreas y, en el caso de las colecciones biológicas, tiene una relevancia significativa. Mientras que la fecha de colecta y la taxonomía responden a las preguntas *cuándo* y *qué* respectivamente, la geografía nos permite conocer la respuesta a *dónde* ocurrió la observación.

La disponibilidad de información sobre la distribución geográfica de las especies se ha visto incrementada rápidamente con la digitalización de colecciones biológicas de museos y herbarios, lo cual ha permitido que estos recursos sean utilizados ampliamente en diversas áreas de investigación, facilitando la comprensión de los patrones y procesos de biodiversidad (Zizka *et al.*, 2019).

De acuerdo con Dalcin (2005), los datos espaciales son aquellos elementos de datos que representan la posición geográfica de la observación o el evento de recolección, lo cual incluye nombres de países, estados, municipios o localidades y coordenadas.

Sin embargo, la información geográfica proveniente de los ejemplares biológicos puede presentar diversos formatos en la nomenclatura de los lugares referidos, así como errores de distinta índole. Si existen problemas en la calidad de los datos, pueden hacer que disminuya su utilidad y certeza (Zizka *et al.*, 2019). Es por este motivo que llevar a cabo el control de calidad del dominio geografía resulta esencial para la fiabilidad de los datos.

Estándares de datos

Los campos del dominio de geografía siguen las recomendaciones estipuladas por el estándar Darwin Core. En particular, para el campo *countryCode* (código de país), que alberga el código estándar correspondiente al país del registro (DGRU, 2019), el estándar Darwin Core sugiere que la mejor práctica es el uso de un código de país ISO 3166-1-alpha-2 (Darwin Core Mantainance Group, 2021). El objetivo en general de los códigos ISO 3166 es proporcionar códigos internacionalmente reconocidos de letras y números que se pueden utilizar para referirse a países y sus subdivisiones (International Organization for Standarization [ISO], s.f.). El uso de códigos de país reduce la probabilidad de errores debidos a variaciones en los nombres de los países en diferentes idiomas, ya que una combinación de letras exclusiva para cada país es comprensible en todo mundo, independientemente del idioma utilizado (ISO, s.f.). De manera específica, el código alpha-2 es un código de dos letras que representa el nombre de un país y está recomendado como un código de propósito general (ISO, s.f.). Por ejemplo, para el caso de México el código que le corresponde es "MX", mientras que, para Estados Unidos, es "US".

Base de datos

El estándar Darwin Core ofrece un conjunto de campos clave que desempeñan un papel esencial en el control de calidad de datos geográficos en contextos biológicos. Estos campos, que incluyen *higherGeography*, *continent*, *waterBody*, *islandGroup*, *island*, *country*, *countryCode*, *stateProvince* y *county*, se utilizan para describir la ubicación geográfica de las observaciones registradas en las colecciones biológicas y proporcionan una estructura estandarizada para capturar información sobre la geografía de manera detallada y precisa. La correcta implementación y validación de estos campos en los conjuntos de datos garantiza que la

información geográfica sea confiable. En la siguiente tabla, se listan estos campos, así como los campos de control e identificadores recomendados para desarrollar la base de datos de este dominio. Además, se indica si su uso es obligatorio u opcional.

Cuadro 10. Campos empleados en el control de calidad del dominio geografía

Tipo de campo	*Nombre del campo*	*Estatus (obligatorio, opcional)*
Campos de registro	*higherGeography*	Opcional
	continent	Opcional
	waterBody	Opcional
	islandGroup	Opcional
	island	Opcional
	country	Obligatorio
	countryCode	Obligatorio
	stateProvince	Opcional
	county	Opcional
Campos de control	lastModified	Obligatorio
	lastModifiedUser	Obligatorio
Identificadores	occurrenceID	Obligatorio
	uuid	Obligatorio
	datasetID	Obligatorio

Nota. En todos los "campos de registro" en la base de datos se incluyen campos con el mismo nombre que el campo de registro, más un sufijo para indicar los datos originales (sufijo "_o"), las calificaciones (con sufijo "_qi") y sus permisos o banderas (con sufijo "_pub").

Fuente: Elaboración propia con base en *Darwin Core Quick Reference Guide* (Grupo de Mantenimiento Darwin Core, 2021).

Catálogos

En el dominio de geografía, los catálogos cumplen distintas funciones, entre las cuales se encuentran:

- Estandarización de los nombres de los lugares referidos, evitando variaciones innecesarias.
- Definición del nivel administrativo al que pertenece la información, determinando si se trata de un país, estado o municipio.
- Evaluación de la consistencia del dato.

Para llevar a cabo un riguroso control de calidad de los datos geográficos, se realiza una comparación de estos con catálogos provenientes de instituciones especializadas y fuentes confiables. Estos catálogos pueden ser obtenidos de tres formas:

- Descarga directa de un catálogo que contenga la información necesaria, como nombres de poblaciones, islas y reservas naturales.

- Descarga de una capa de datos espaciales que abarque los temas y la cobertura geográfica necesarios para el control de calidad y, a partir de los atributos asociados a dicha capa, se extrae el catálogo requerido.
- Utilización de un Servicio de Mapas Web (WMS, por sus siglas en inglés), un estándar ampliamente utilizado para compartir mapas en internet en formatos de imagen como PNG, GIF o JPEG. Un servicio WMS se utiliza para acceder a información cartográfica a través de Internet. Los SIG en computadoras de escritorio son comunes para acceder a estos servicios (INEGI, s.f.).

En la segunda forma, se agrega como catálogo la Base de Datos de las Áreas Administrativas Globales (GADM, por sus siglas en inglés), la cual provee mapas e información espacial de las áreas administrativas mundiales, es decir, de las fronteras de los países y otras subdivisiones menores (Global Administrative Areas, 2023). Asimismo, incluye información acerca de los atributos para cada área, como el nombre del área en cuestión, variantes del nombre, denominación del nivel administrativo, código de país en distintos formatos, entre otros datos.

La versión 2.8 de este catálogo se utiliza para el control de calidad del campo *country*, ya que cuenta con un formato estandarizado de los nombres de los distintos países en español. Una desventaja de este catálogo es que en versiones recientes no incluye el nombre de los países en español, lo que hace necesario recurrir a otras fuentes para poder realizar el control de calidad de estos datos. Como segunda fuente para el control de calidad de dicho campo, se utiliza la base de datos geográfica GeoNames, la cual integra información geográfica de nombres de lugares en diversos lenguajes, elevaciones, coordenadas, entre otra información (GeoNames, s.f.). Por otra parte, el catálogo GADM versión 2.8 cuenta con los códigos de país en formato ISO 3166-1-alpha-2, por lo que es utilizado para el control de calidad del campo *countryCode*.

Para el control de calidad de los campos *stateProvince* y *county* en ejemplares pertenecientes a países distintos de México, se utiliza la versión más reciente del catálogo GADM, debido a que este cuenta con el formato más actual de los nombres de las áreas administrativas mundiales en distintos niveles.

En el caso de los ejemplares pertenecientes a México, los catálogos que se usan de forma prioritaria son los generados por el Instituto Nacional de Estadística y Geografía (INEGI), una institución especializada en la generación de información estadística y geográfica del territorio mexicano. El control de calidad de los campos *stateProvince* y *county* se realiza con base en el Marco Geoestadístico (MG), que proporciona capas de datos espaciales en formato vectorial generadas por dicha institución. El MG cuenta con un formato de descarga que incluye las capas de datos espaciales de las divisiones administrativas que componen al territorio nacional en formato Shapefile. A partir de los atributos asociados a las capas, se obtienen los nombres de los 32 estados y sus respectivos municipios, los cuales son utilizados para el control de calidad.

La información de cuerpos de agua (océanos, lagos, lagunas, etc.), islas y grupos de islas se registran en los campos *waterBody*, *island* e *islandGroup*, respectivamente. Para el control de calidad de los datos nacionales de estos campos se utilizan principalmente los WMS. A continuación, se enlistan algunos de los catálogos que se utilizan para el control de calidad de las colectas nacionales (consulte el Anexo I para obtener la lista completa de temas utilizados):

- Información topográfica 1:50,000
    - Cuerpos de agua
    - Textos de cuerpos de agua
- Territorio insular
    - Arrecifes
    - Oceánicos costeros
    - Terrestres (islas en cuerpos de agua continentales)- 2013

Además de los WMS, para estos campos se utiliza la información proveniente de las siguientes capas de datos espaciales del INEGI:

- Archipiélagos
- Conjuntos de datos vectoriales de información topográfica escala 1:250 000 por Entidad Federativa serie VI.
    - Cuerpos de agua
    - Elementos insulares

Adicionalmente al INEGI, otra institución cuya información es confiable para el control de calidad de los datos geográficos nacionales es la CONABIO. De esta agencia gubernamental, se utilizan los datos asociados a la capa de Regiones Marinas (CONABIO, 2017) para colectas nacionales realizadas en el territorio marítimo de México.

Cuando no se encuentran datos geográficos naciones en los productos del INEGI o se trata de datos de otros países, se recurre a las siguientes fuentes:

- Base de datos geográfica GeoNames
- Google Maps
- Base de datos de áreas administrativas globales GADM
- Bibliografía especializada proveniente de fuentes confiables; en ocasiones, se cuenta con bibliografía generada a partir de investigaciones en las que se mencionan los datos geográficos de colecta, lo que permite corroborar cierta información. En otros casos, se recurre a referencias académicas de fuentes confiables.

Cuadro 11. Catálogos utilizados para los campos del dominio geografía

Campo	*Catálogo*	*Observaciones*
higherGeography	NA	Se asignan los datos que resulten consistentes a partir del control de calidad de los campos *continent,*

Campo	*Catálogo*	*Observaciones*
		waterBody, islandGroup, island, country, stateProvince y *county*.
continent	NA	Asignado a partir del valor consistente de *country*. Se utiliza el vocabulario controlado designado para este campo.
waterBody	WMS INEGI: Información topográfica 1:50,000 Cuerpos de agua Textos de cuerpos de agua Regiones marinas (CONABIO, 2017) GeoNames Bibliografía especializada (a partir de la investigación del analista o proporcionado por el proveedor de datos)	El catálogo utilizado depende de la información que contenga el campo.
islandGroup	WMS INEGI: Territorio insular Arrecifes Oceánicos costeros Conjunto de datos del Territorio Insular Mexicano. Escala 1:50,000. Versión 2.0 GeoNames Bibliografía especializada (a partir de la investigación del analista o proporcionado por el proveedor de datos)	El catálogo utilizado depende de la información que contenga el campo.
island	Conjunto de datos del Territorio Insular Mexicano. Escala 1:50,000. Versión 2.0 GeoNames Bibliografía especializada (a partir de la investigación del analista o proporcionado por el proveedor de datos)	El catálogo utilizado depende de la información que contenga el campo.
country	GADM versión 2.8 GeoNames	GADM es prioritario, si no se encuentra la información en este catálogo, entonces se recurre a GeoNames.
countryCode	GADM versión 2.8: ISO 3166 - 1 - alpha-2	
stateProvince	Registros nacionales: MG, INEGI Registros internacionales:	

Campo	*Catálogo*	*Observaciones*
	GADM GeoNames	
county	Registros nacionales: MG, INEGI Registros internacionales: GADM GeoNames	

Fuente: Elaboración propia.

Se recomienda que, en la medida de lo posible, dichos catálogos se encuentren en el mismo formato de la base de datos para poder utilizar las herramientas disponibles y permitir la comparación entre los datos de la base y el catálogo.

Vocabularios controlados

Los vocabularios controlados desempeñan un papel fundamental en el control de calidad de bases de datos geográficos de las colecciones biológicas, ya que ayudan a asegurar la consistencia y precisión en la descripción de las ubicaciones referidas. Al utilizar términos estandarizados, se evitan equivocaciones y se facilita la comparación y el análisis de datos.

Los campos *country*, *countryCode*, *stateProvince* y *county* deben seguir los vocabularios controlados definidos por los catálogos antes mencionados. Además, el campo *continent* depende directamente del campo *country*, pues este define el continente al que pertenece el registro en cuestión. Los continentes disponibles en el vocabulario controlado África, América, Asia, Europa y Oceanía.

Herramientas de datos

Es recomendable realizar un análisis exploratorio de la base de datos para identificar campos con datos, aquellos sin datos, y cualquier variación en el formato de la misma información. Esto permitirá establecer las estrategias para el procesamiento de la información, pues las características de los datos, así como sus problemáticas, cambian en cada colección.

1. Comparación con catálogos
   Para el control de calidad en el dominio de geografía, se lleva a cabo una comparación de los datos en los campos con los catálogos mencionados.
   a. *country*.
      i. El primer campo que es comparado es el de *country*, pues es a partir de este que podremos definir cuáles son los catálogos que aplicarán para el control de

calidad de los campos *stateProvince* y *county*, así como la asignación de *countryCode*.

ii. El campo *country* se compara con el catálogo GADM versión 2.8 utilizando el campo de nombre de país en español. Aquellos registros que son iguales al catálogo son calificados como consistentes. Los registros que no coinciden con el catálogo son detectados y se verifica si el dato registrado tiene algún tipo de variante en el nombre, ya sea por el formato (por ejemplo "República Mexicana") o porque existe algún error de digitación u ortográfico (por ejemplo, "Mexico"). Si es el caso, se estandariza el dato a formato que indica el catálogo (siguiendo el ejemplo anterior, el valor procesado de acuerdo con el catálogo GADM sería "México").

iii. Si continúan datos del campo *country* sin coincidir con los catálogos o el campo es nulo, estos registros se quedan como pendientes para revisión posterior, esperando que en pasos subsecuentes del control de calidad se pueda dilucidar el valor que le corresponde con base en el resto de datos geográficos del registro. Si persisten datos que no coinciden con los catálogos, se califican como inconsistentes.

b. *countryCode*

i. Una vez que se tiene certeza de que el dato en el campo *country* es consistente (ya sea de origen o por modificaciones en el proceso de control de calidad), el valor de *countryCode* es asignado a partir del catálogo GADM 2.8. Se realiza una consulta de comparación entre el campo *country* de la base de datos y el campo que alberga los nombres de países en el catálogo, asignando el valor que le corresponda en el campo *countryCode*.

c. *stateProvince* y *county*

i. Luego, con base en el campo *country*, se determina el catálogo que se utilizará para el control de calidad de *stateProvince* y *county*. Si se trata de datos pertenecientes a México, el catálogo que será utilizado es el Marco Geoestadístico de INEGI en su versión más reciente. Si los datos no pertenecen a México, deberá utilizarse el catálogo GADM, en la versión más actualizada. A continuación, se comparan los campos *stateProvince* y *county* con los catálogos que les correspondan.

- En el caso de los ejemplares nacionales, los datos del campo *stateProvince* se comparan con el campo del MG de INEGI que corresponda con este nivel administrativo.
- Al mismo tiempo, si se cuenta con datos en el campo *county*, estos son comparados con su nivel correspondiente en dicho catálogo.
- Estas acciones se realizan con el catálogo GADM para aquellos datos que no pertenecen a México.

ii. Los datos que coinciden con los valores del catálogo se consideran consistentes, mientras que aquellos que no coinciden son evaluados para detectar si existe algún tipo de problemática que impida su coincidencia con los valores del catálogo (tal como se llevó a cabo con el campo *country*) y, una

vez localizado el error, se realizan las modificaciones correspondientes. Si persisten datos que no coincidan con los catálogos deberán ser calificados como inconsistentes.

d. *waterBody*, *island*, *islandGroup*
    i. La comparación entre los datos de los campos *waterBody*, *island* e *islandGroup* y los catálogos tiene múltiples finalidades, como la detección de errores ortográficos (por ejemplo, *island_o*: Isla Tiburon, *island*: Isla Tiburón), estandarización de los datos (como, *island_o*: Waigeu, Isl, *island*: Waigeo Island) y verificación de la consistencia de los datos reportados con la geografía superior revisada hasta el momento. Por ejemplo, el campo *waterBody* se procesa con base en los atributos asociados a la capa de "Regiones Marinas Prioritarias de México" (CONABIO, 2001) para la Colección Diatomeas (Bacillariophyceae) y Dinoflageladas (Dinophyceae), que son planctónicas y frecuentes en la región sur del Golfo de México. Dicha capa contiene las principales regiones marinas nacionales, incluidas las que se referían al campo *waterBody* de dicha colección (Golfo de México y Mar Caribe). Los datos contenidos en el campo fueron comparados con los indicados en el catálogo, estandarizando los nombres de los cuerpos de agua en los casos que fuera necesario, permitiendo la validación, corrección o asignación de datos.

e. *higherGeography*
    i. El campo *higherGeography* suele ser nulo de origen, sin embargo, se compone de la concatenación de los datos consistentes de los campos *continent*, *waterBody*, *islandGroup*, *island*, *country*, *stateProvince* y *county*, separados por la barra vertical (|), como lo indican las recomendaciones para el campo del estándar Darwin Core. Por ejemplo, "América | México | Ciudad de México | Coyoacán". Este campo se asigna como parte del último paso del control de calidad de los datos geográficos en su totalidad, pues solo podrán formar parte del concatenado aquellos datos que sean consistentes.

2. Asignación de datos

En algunos casos, los registros pueden carecer de datos en alguno de los campos. Por ejemplo, pueden presentar datos en *country*, *county* y *locality*, pero no poseer datos en *stateProvince*. Sin embargo, la información disponible en el resto de los campos permite que el analista sea capaz de asignar el valor faltante. Por ejemplo:

- *country*: México
- *stateProvince*: sin dato
- *county*: Coyoacán
- *locality*: Ciudad Universitaria, UNAM

En este caso, con base en los catálogos aprobados y la información disponible en los demás campos geográficos, se puede realizar la asignación de "Ciudad de México" como el estado correspondiente.

Cabe resaltar que este procedimiento debe llevarse a cabo de forma cuidadosa, ya que una asignación con información errónea no solo no enriquecería la información geográfica, sino que también podrían generar errores, disminuyendo la fiabilidad del registro.

En la Figura 6 se muestra el procedimiento seguido en el control de calidad de los datos que albergan los campos del dominio geografía.

Evaluación de la calidad de datos

La evaluación de la calidad de datos a través del uso de calificaciones es un proceso importante en el control de calidad. Este proceso permite determinar el nivel de confiabilidad de los datos geográficos y, por ende, su utilidad para la toma de decisiones y análisis. Las calificaciones permiten asignar un valor numérico a diversos aspectos de los datos. Al calificar los datos, se puede informar al proveedor de datos sobre qué información ha sido asignada durante el procesamiento de los datos, si los nombres de las referencias geográficas han sido estandarizados con base en catálogos o si los datos han sido eliminados o reasignados a otros campos. En resumen, estas calificaciones informan al proveedor de datos sobre el estado actual de su colección, así como de aquellas modificaciones realizadas durante el control de calidad. El Cuadro 12 desarrolla las calificaciones utilizadas en el control de calidad del dominio geografía:

Cuadro 12. Calificaciones empleadas en la evaluación de la calidad de los datos del dominio geografía

	Calificación (qi)	*Situación*	*Ejemplos*
consistente	1	Consistente de origen: el dato original y procesado son iguales.	*country_o*: México *country*: México
	0.9	Modificado por variantes de escritura (por ejemplo, errores de digitación, ortografía, cambio de minúsculas o mayúsculas, inserciones o espacios). La identidad del dato se mantiene.	*country_o*: MEXICO *country*: México
	0.8	Dato asignado durante el proceso de control de calidad, es decir, no está presente en el campo original.	*countryCode_o*: sin dato *countryCode*: MX
	0.7	Modificado por estandarización con respecto al catálogo.	*stateProvince_o*: Distrito Federal *stateProvince*: Ciudad de México El valor fue modificado ya que el catálogo utilizado reconoce que para dicha entidad el nombre válido es Ciudad de México.
inconsistente	0.4	Requiere la resolución del proveedor de datos tras realizar la revisión en el control de calidad.	*locality_o*: Entró a México en cargamento de árboles de Navidad por Nuevo Laredo, Tamaulipas. *country_o*: Estados Unidos *country_qi*: 0.4 En este caso, se debía verificar con el proveedor de datos cuáles debían ser los datos geográficos adecuados para el

			ejemplar, los indicados en *locality* o el valor en *country*.
	0.1	No fue posible encontrar relación lógica con la definición del campo según el estándar, y no se ha encontrado confirmación en ningún catálogo o referencia, lo que resulta inconsistente.	*country*: México *stateProvince*: Guanajuato *county_o*: Jaral *county*: Jaral "Jaral" no está reportado en el MG de INEGI como municipio del estado de Guanajuato, por lo que es un dato inconsistente.
Nulo	0	El valor original se eliminó del campo.	*county_o*: Distrito de Ixtlán *county*: sin dato
	Vacío (sin dato)	Cuando no existe valor para revisar	*island_o*: sin dato *island*: sin dato

Nota. Las calificaciones se dividen en tres grupos: consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Perfil del personal

El desarrollo del perfil del personal encargado de la calidad de los datos geográficos es esencial para garantizar la fiabilidad y utilidad de la información. Es necesario que el profesional posea diversos atributos y habilidades para asegurar que los datos geográficos sean precisos y coherentes, lo que a su vez respalda una toma de decisiones basada en datos confiables. A continuación, se detallan las características básicas del personal en este contexto:

1. Conocimientos en geografía
   Es importante que la persona comprenda conceptos geográficos generales, como niveles administrativos (por ejemplo, país, estado, municipio), sistemas de coordenadas, proyecciones cartográficas, etc. Esto permite una mejor comprensión de la naturaleza de los datos geográficos y evaluar su calidad.
2. Familiaridad con los SIG
   Debe tener experiencia en el uso de software y herramientas SIG, como ArcGIS, QGIS u otros. Esto le permitirá trabajar con capas geoespaciales, realizar análisis espaciales y validar la calidad de los datos utilizando las herramientas disponibles en dichos programas.
3. Manejo de catálogos geográficos
   La persona deberá de ser capaz de utilizar catálogos geográficos provenientes de fuentes fidedignas que le permitan realizar el control de calidad, pues es con base en estos que evaluará la consistencia de los datos.
4. Conocimientos en bases de datos espaciales
   Es importante que la persona esté familiarizada con el manejo de bases de datos espaciales. Deben entender cómo almacenar, consultar y manipular datos geográficos en estas bases, así como realizar validaciones y correcciones de calidad de los datos geográficos almacenados.

5. Atención al detalle
   El control de calidad de datos geográficos requiere una gran atención al detalle para identificar problemas y anomalías en los datos. La persona debe ser minuciosa en la revisión de los datos y tener la capacidad de detectar errores o inconsistencias.
6. Capacidad analítica
   Debe tener habilidades analíticas para evaluar la calidad de los datos geográficos y comprender los posibles impactos de los errores en los análisis y aplicaciones que utilizan esos datos.
7. Comunicación y colaboración
   La persona debe ser capaz de comunicarse efectivamente con otros miembros del equipo, usuarios finales y partes interesadas, a fin de entender los requisitos, informar sobre problemas de calidad y colaborar en la resolución de los mismos.
8. Curiosidad y aprendizaje continuo
   Dado que el campo de los datos geográficos y las tecnologías asociadas está en constante evolución, es importante que la persona tenga una actitud curiosa y esté dispuesta a aprender nuevas técnicas, herramientas y estándares relacionados con la calidad de los datos geográficos.

Asignación de roles

En el dominio de geografía, se requiere la colaboración del coordinador y el analista en metodología y estadística. Las actividades particulares para el control de calidad de este dominio las llevan a cabo un analista líder en datos geográficos y un analista de datos. A continuación, se enuncian dichas actividades:

1. Coordinador
2. Analista en metodología y estadística
3. Analista líder de datos geográficos
   Campos de los que es responsable: *higherGeography*, *continent*, *country*, *countryCode*, *stateProvince*, *county*
   - Evaluación de las tareas realizadas por el analista de datos.
   - Procesamiento de los datos que quedaron sin una calificación consistente (evalúa casos especiales, por ejemplo, si presentan alguna variante o si el nombre de país, estado o municipio requiere ser actualizado).
   - Detección y calificación de datos inconsistentes.
   - Asignación de datos faltantes (por ejemplo, casos en los que el registro no cuenta con datos en cierto campo y este puede ser asignado a partir de datos presentes en otros campos).
   - Procesamiento (asignación o validación) de los campos *island*, *islandGroup*, *waterBody*.
   - Asignación de datos en el campo *continent* que quedaron pendientes.

- Asignación de *higherGeography*.

4. Analista de datos

Campos de los que es responsable: *country*, *countryCode*, *stateProvince*, *county*

- Limpieza de datos.
- Comparación de los campos bajo su responsabilidad con catálogos.
- Calificación datos consistentes (aquellos que coincidieron con los catálogos).
- Estandarización de los datos.
- Notificación de casos particulares y errores detectados.
- Asignación datos en el campo *continent* con base en los datos de *country* que resultaron consistentes.

Coordenadas

Las colectas biológicas que cuentan con coordenadas son de gran importancia, ya que brindan información sobre la distribución geográfica de las especies. Entender los patrones geográficos de las especies es fundamental para conocer el estado actual de la biodiversidad, además de dar orientación en la toma decisiones en torno a su conservación (Jetz *et al*., 2019, como se citó en Arlé *et al*., 2021). Si bien las bases de datos biológicos georreferenciados han aumentado rápidamente en las últimas décadas, la inexactitud de las mismas ocurre también de forma recurrente por problemas en los procesos de georreferenciación e identificaciones erróneas, lo cual resulta en información geográfica incorrecta. Esta situación es común en bases de datos biológicas. Si estos errores no son detectados, las inexactitudes pueden distorsionar los resultados de los análisis tales como el modelado de la distribución de especies (Arlé *et al*., 2021).

Es por este motivo que el control de calidad de las coordenadas de los registros en las colecciones biológicas es de gran importancia.

Estándares de datos

En el dominio de las coordenadas, se siguen las recomendaciones estipuladas por el estándar Darwin Core para los campos *verbatimLatitude*, *verbatimLongitude*, *decimalLatitude* y *decimalLongitude*. El primer par de campos permite la preservación de los datos originales, ya que aquí se conservan las coordenadas tal como se capturaron originalmente, manteniendo un registro auténtico de los datos. Por otra parte, los campos *decimalLatitude* y *decimalLongitude* ofrecen una representación estandarizada y estructurada de las coordenadas geográficas en formato decimal. Esto asegura la coherencia en los datos y facilita la visualización, comparación y análisis en SIG u otras herramientas geoespaciales. Finalmente, el uso de estos campos estandarizados en el dominio de coordenadas promueve la interoperabilidad entre diferentes bases de datos.

Base de datos

Para la creación de la base de datos del dominio de coordenadas se utilizan los campos *decimalLatitude* y *decimalLongitude*, que representan las coordenadas geográficas en formato decimal, y *verbatimLatitude* y *verbatimLongitude,* en los cuales son registradas las coordenadas tal como se obtuvieron originalmente. Su adecuada aplicación garantiza que los datos sean coherentes y fiables. A continuación, se muestran los campos descritos previamente, además de los identificadores y los campos dedicados al control de cambios. Además, se indica si son de carácter obligatorio u opcional.

Cuadro 13. Campos empleados en el control de calidad del dominio coordenadas

Tipo de campo	*Nombre del campo*	*Estatus (obligatorio, opcional)*
Campos de registro	*decimalLatitude*	Opcional
	decimalLongitude	Opcional
	verbatimLatitude	Opcional
	verbatimLongitude	Opcional
Campos de control	lastModified	Obligatorio
	lastModifiedUser	Obligatorio
Identificadores	occurrenceID	Obligatorio
	uuid	Obligatorio
	datasetID	Obligatorio

Nota. En todos los "campos de registro" en la base de datos se incluyen campos con el mismo nombre que el campo de registro, más un sufijo para indicar los datos originales (sufijo "_o"), las calificaciones (con sufijo "_qi") y sus permisos o banderas (con sufijo "_pub").

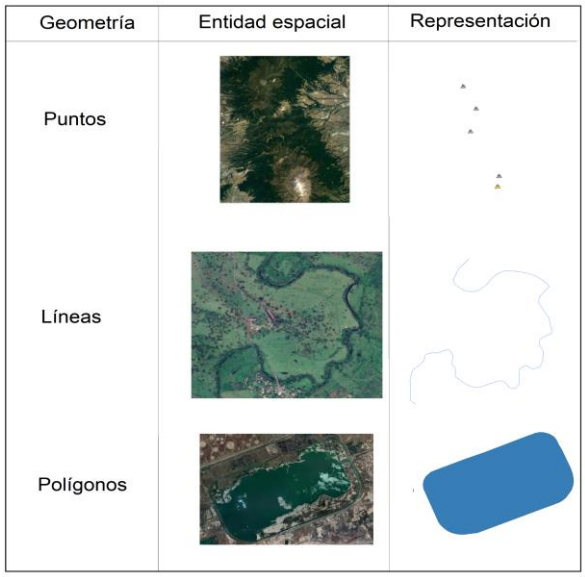
Fuente: Elaboración propia con base en *Darwin Core Quick Reference Guide* (Grupo de Mantenimiento Darwin Core, 2021).

Catálogos

Para el control de calidad de las coordenadas, se utilizan como catálogos que se basan principalmente en capas de datos espaciales provenientes de fuentes confiables. Estas capas se proyectan en alguno de los diversos SIG disponibles, haciendo posible determinar si las coordenadas son consistentes con la información indicada en los demás campos que albergan los datos geográficos. De forma secundaria, se utilizan WMS y, en última instancia, bibliografía especializada que proporciona información sobre la ubicación del dato geográfico referido.

Las capas de datos espaciales utilizadas en el control de calidad de los datos geográficos son principalmente datos vectoriales. Estos albergan objetos espaciales, en los que se representan entidades del mundo real, como un volcán, un río o una población mediante geometrías de tipo punto, línea o polígono (QGIS project, 2022).

Figura 6. Representación de entidades espaciales como geometrías en las capas de datos espaciales de tipo vectorial



Fuente: Elaboración propia con imágenes satelitales de Google Earth.

En el contexto del dominio de geografía, el catálogo elegido para el control de calidad de las coordenadas dependerá de si se trata de datos nacionales o internacionales. Para los datos nacionales, se dará prioridad a la información proveniente del INEGI y otras instituciones nacionales especializadas, mientras que en el caso de datos internacionales, serán comparados principalmente con el catálogo GADM.

Si la información necesaria no se encuentra en estos catálogos, entonces se recurre a sitios como GeoNames o Google Maps. Asimismo, se utilizan capas geoespaciales provenientes de organizaciones calificadas que contengan los datos requeridos.

A continuación, se mencionan los catálogos más utilizados en el control de calidad de las coordenadas.

Cuadro 14. Catálogos utilizados en el control de calidad del dominio coordenadas

Catálogo	*Temas utilizados*	*Formato*	*Campos comparados*	*Aplicación*
MG INEGI	Áreas geoestadísticas estatales y municipales. Polígonos de localidades urbanas y rurales.	Shapefile	*stateProvince, county, locality*	Registros nacionales
GADM	Niveles administrativos 0 al 5	GeoPackage	*country, countryCode, stateProvince, county, locality*	Registros internacionales

WMS INEGI	Datos del relieve Información topográfica 1:50,000 Marco Geoestadístico Recursos naturales Red nacional de caminos Registro de nombres geográficos Territorio insular Zonas hidrogeológicas	Servicio de mapas Web (WMS)	*locality*	Registros nacionales
Conjunto de datos del Territorio Insular Mexicano. Escala 1:50,000. Versión 2.0	NA	Shapefile	*island, islandGroup*	Registros nacionales
GeoNames	NA	Base de datos geográfica (acceso a través de internet)	*country, locality*	Registros nacionales e internacionales
Google Maps	NA	Servidor de aplicaciones de mapas web	*locality*	Registros nacionales e internacionales

Fuente: Elaboración propia.

Como parte de la investigación realizada para el control de calidad, es posible identificar sitios y sus coordenadas asociadas que han sido obtenidas de fuentes especializadas, por lo que es de gran utilidad contar con un catálogo interno que compile la ubicación precisa de los rasgos geográficos encontrados, junto con la fuente de referencia.

Vocabularios controlados

En el dominio de las coordenadas, los vocabularios controlados establecen los elementos necesarios para la normalización de las coordenadas en formato original, registradas en los campos *verbatimLatitude* y *verbatimLongitude*, lo que facilita su posterior conversión al formato requerido por el estándar Darwin Core, en los campos *decimalLatitude* y *decimalLongitude*. Además, en este dominio se ha observado que el uso de vocabularios controlados permite detectar de manera temprana los posibles errores de origen. A continuación, se describen los vocabularios controlados utilizados en los campos *verbatimLatitude*, *verbatimLongitude*, *decimalLatitude* y *decimalLongitude*.

1. *verbatimLatitude* y *verbatimLongitude*

Los datos en estos campos usualmente se presentan en coordenadas sexagesimales (grados, minutos y segundos) y requieren su posterior conversión a grados decimales, para lo cual deben cumplir con un formato específico.

Para ambos campos, los grados deben estar representados por el símbolo "°", los minutos por una comilla simple "'" y los segundos por dos comillas simples "''". Los minutos y segundos deben contener números entre 0 y 59. De manera particular los datos de cada campo deben ubicarse dentro de un intervalo de grados y una dirección, determinados por su posición con respecto a los hemisferios norte/sur y este/oeste. Ejemplo:

20° 32' 50'' N
88° 49' 38'' W

verbatimLatitude

Para registros cuyo país se ubique al norte del ecuador, se debe indicar "N" al final de los valores numéricos, y "S" para los que se ubican al sur del ecuador. El intervalo numérico va de 0 a 90 grados. Por ejemplo, la siguiente coordenada pertenece a México, por lo que le corresponde una dirección norte:

20° 23' 9'' N

verbatimLongitude

Si el país indicado en el registro se ubica al este del Meridiano de Greenwich, se debe indicar "E", y si se encuentra al oeste, "W". El intervalo numérico va de 0 a 180 grados. Siguiendo el ejemplo anterior, la dirección que le corresponde es W, ya que se encuentra al oeste del meridiano de referencia:

98° 55' 30'' W

2. *decimalLatitude* y *decimalLongitude*

En estos campos, ambos datos deberán presentar un formato de coordenadas geográficas en grados decimales (por ejemplo, 21.23303; -89.89217).

decimalLatitude

Para registros cuyos países se ubiquen al norte del ecuador, deberán de presentar valores positivos, mientras que los que están al sur, valores negativos. El intervalo numérico de los datos es de -90 a 90 grados. Continuando con el ejemplo previo, dichas coordenadas sexagesimales, al ser transformadas a grados decimales, presentan el siguiente valor para *decimalLatitude*:

20.385833

decimalLongitude

Valores positivos para los registros que están al este del Meridiano de Greenwich y los negativos están al oeste. El intervalo numérico de los datos es de -180 a 180 grados. En este caso, el valor de *decimalLongitude* del ejemplo sería el siguiente:

-98.925

Herramientas de datos

El control de calidad de las coordenadas no se limita a los campos que las albergan, sino que es un proceso que debe contemplar toda la información geográfica, ya que su consistencia está en función del contexto que brindan el resto de los campos.

Los sistemas de coordenadas utilizados comúnmente en las colecciones biológicas están basados en coordenadas geográficas, es decir, latitud y longitud (Chapman & Wieczorek, 2020). Las coordenadas geográficas son una manera conveniente para definir una localidad de forma más específica en comparación con una simple descripción. Además, a partir de estos datos espaciales, se puede visualizar la ubicación de las colectas y realizar análisis en SIG.

Las coordenadas geográficas se pueden presentar en diversos formatos, como grados decimales, grados-minutos-segundos o grados-minutos decimales, siendo los grados decimales los más utilizados (Chapman & Wieczorek, 2020), y son el formato requerido por el estándar Darwin Core para los campos *decimalLatitude* y *decimalLongitude*.

El control de calidad del dominio de coordenadas se realiza de la siguiente manera:

1. Limpieza de los datos

   Inicialmente se realiza una limpieza de los datos, verificando que no presenten espacios, errores ortotipográficos, etc.

2. Homogeneización de datos

   Los campos *verbatimLatitude* y *verbatimLongitude* pueden presentar diversos formatos en las coordenadas. Por ejemplo, se han observado casos en que los símbolos de las coordenadas estaban sustituidos por comas, espacios, comillas o tildes, por lo que son modificados agregando los símbolos correspondientes, por ejemplo, 18°13´7.6" se convierte en 18°13’7.6**’’ y** 18**,**9**,**50**,** se convierte en 18°9’50**’’.**

3. Reasignación de datos

   Existen casos en los que hay datos que no corresponden con la información que debe contener el campo, por lo que son reasignados al campo adecuado. Por ejemplo, que el campo de *verbatimLatitude* indique el nombre de un país, en este caso el valor deberá de ser reasignado al campo *country*.

4. Eliminación de datos

   Se realiza la eliminación de datos que no se pueden asignar y no son de carácter informativo para el registro en cuestión. Por ejemplo, la representación de datos nulos con valores, como 99999 9999 9999, @@@@, ND y SD, pues no son informativos y deberán ser eliminados.

5. Verificación de la definición del campo

   Después de la limpieza, estandarización y eliminación de datos, es importante determinar si los datos de las coordenadas cumplen con los requisitos del estándar Darwin Core:

   a. Datos presentes en ambos campos

Las coordenadas se conforman por dos datos, latitud y longitud, que dependen mutuamente para indicar la ubicación de un evento. Si uno de los campos no contiene valor, se vuelve imposible determinar la ubicación. En este caso, el campo que sí contiene datos se califica como inconsistente, mientras que el otro permanece nulo.

b. Correcta asignación de datos

Con frecuencia, se detecta que los datos de latitud han sido asignados incorrectamente al campo de longitud (o viceversa), por lo que estos no pueden ser procesados correctamente. En los ejemplos siguientes, se refieren a registros de ubicaciones en México, donde se espera que la latitud esté presentada con valores positivos o una dirección al norte, y la longitud con valores negativos o una dirección hacia el oeste:

decimalLatitude: -96.542814

decimalLongitude: 18.096881

verbatimLatitude: 92°40'N

verbatimLongitude: 16°56'O

En ambos ejemplos es evidente que los datos están asignados al campo contrario. En esta situación, se deberá reasignar el valor al campo que le corresponde.

c. Dirección o los valores (positivo o negativo) en los datos

Es habitual que los valores numéricos de las coordenadas estén asignados correctamente, pero que la dirección asociada sea incorrecta. A continuación, se presenta el caso donde el ejemplar indica pertenecer a México, por lo que le corresponde una latitud norte y una longitud oeste, sin embargo, estos datos se encuentran asignados de forma errónea:

verbatimLatitude: 21°02'53'' W

verbatimLongitude: 98°39'14'' N

En este caso, si bien los valores numéricos son consistentes, las direcciones son incorrectas e impiden su correcto procesamiento. Es importante ubicar estos errores y modificarlos al valor correspondiente.

Otro formato de error es aquél en el que la dirección se representa incorrectamente mediante los símbolos de positivo y negativo (+,-). Por ejemplo:

decimalLatitude: -17.98333

decimalLongitude: -89.38611

En ejemplo, se refiere a un registro perteneciente a México, donde se espera que la latitud se exprese con valores positivos y la longitud con valores negativos. Sin embargo, la latitud contiene un valor negativo, lo que provoca un cambio significativo en la ubicación, pasando a localizarse en algún lugar del Océano Pacífico, cercano a Perú. Este tipo de problemáticas deberá resolverse mediante

la eliminación del signo negativo si no corresponde o agregarlos si es necesario, lo cual se decidirá con base en la posición del país con respecto a los hemisferios. Si el país se ubica en el hemisferio norte, le corresponde un valor positivo, si se encuentra en el hemisferio sur, deberá presentar un signo negativo. De la misma manera, será asignado un valor positivo si se encuentra en el hemisferio oriental y un valor negativo si se encuentra en el hemisferio occidental.

d. Cumplimiento de los límites de los datos geográficos

Las coordenadas tienen límites definidos. En el caso de la latitud, los valores positivos están al norte del ecuador, mientras que los negativos están al sur, y el rango numérico válido se encuentra entre -90 y 90. En la longitud los valores positivos están al este del Meridiano de Greenwich y los negativos están al oeste del mismo, y el rango numérico permitido se sitúa entre -180 a 180. Estos intervalos están definidos con base en la ubicación del país. Los campos *verbatimLatitude* y *verbatimLongitude* albergan las coordenadas en el formato original, que habitualmente es sexagesimal y en menor proporción como coordenadas UTM. En el primer formato, se han encontrado errores como el siguiente:

verbatimLatitude: 19°02'320"

verbatimLongitude: 98°18'191"

En este ejemplo, los segundos exceden el límite establecido, que va de 0 a 59.999999, por lo que son inconsistentes.

6. Transformación del formato de las coordenadas

a. Para el procesamiento de las coordenadas en un SIG y con base en lo indicado por el estándar Darwin Core, se requiere que los valores estén en formato de grados decimales; sin embargo, en la mayoría de los casos las coordenadas de los ejemplares biológicos se encuentran en formato sexagesimal y no en grados decimales. En este caso, se deberán transformar al formato requerido. Esto se puede llevar a cabo mediante distintos métodos, sin embargo, se sugiere el uso de una herramienta de la página web *Canadensys*, (Canadensys, s.f.) cuya creación y mantenimiento es responsabilidad de *The Montreal Biodiversity Centre* de Canadá. Esta página fue diseñada para compartir información de especímenes resguardados por las colecciones biológicas de universidades canadienses (Bell *et al.*, 2011). El equipo de Canadensys desarrolló un conjunto de herramientas independientes, entre las cuales se encuentra la Conversión de Coordenadas. Esta herramienta convierte las coordenadas geográficas del formato sexagesimal a grados decimales. Para utilizarla, se deben ingresar en el espacio designado los pares de coordenadas que deberán cumplir los formatos indicados por la página para poder ser procesados, que son esencialmente la homogeneización de datos (Paso 2) y verificar que la dirección en los datos *verbatim* sea consistente (Paso 5c).

Una vez que las coordenadas se someten a la herramienta, el resultado es inmediato, obteniendo las coordenadas en formato de grados decimales. Posteriormente, estas deberán ser asignadas a los campos *decimalLatitude* y *decimalLongitude* y redondeadas a cinco decimales, ya que de esta forma son más precisas (Chapman & Wieczorek, 2020).

Una de las ventajas de esta herramienta es que permite ingresar numerosos pares de coordenadas. Asimismo, resalta las coordenadas que no fueron convertidas, permitiendo detectar si existen errores en la homogeneización o si se trata de un error de origen como los mencionados en el Paso 5d.

La transformación de las coordenadas a un formato de grados decimales, además de cumplir con las especificaciones del estándar Darwin Core, permite que sean proyectadas más fácilmente al SIG.

7. Elección del SIG

   El SIG es una herramienta que permite el análisis y representación de los datos espaciales (Geoinnova, 2021), por lo que es esencial no solo para el control de calidad de las coordenadas, sino también para los datos geográficos contenidos en el resto de los campos. Particularmente, en el caso de las coordenadas, permite su procesamiento mediante el uso de los recursos disponibles en el SIG y la visualización de los catálogos utilizados.

   Existen diversos SIG disponibles, sin embargo, se sugiere el uso de QGIS, un proyecto respaldado por la Open Source Geospatial Foundation (OSGeo). Este software es de código abierto, compatible con Linux, Unix, Mac OSX, Windows y Android, y es capaz de manejar una amplia gama de formatos y capacidades de información geoespacial, abarcando datos vectoriales, datos ráster y conexiones con bases de datos (QGIS, s.f.). Además, cuenta con una gran diversidad de herramientas y recursos utilizados en el control de calidad.

8. Proyección en el SIG

   Una vez que las coordenadas cumplen con el formato de grados decimales, se debe crear un archivo con los datos geográficos del registro. Se recomienda incluir al menos la siguiente información: *country*, *stateProvince*, *county*, y *locality*). Además, este archivo debe contener los identificadores y las coordenadas en grados decimales albergadas en *decimalLatitude* y *decimalLongitude*. Este archivo será proyectado en QGIS para poder compararlo con los catálogos mediante las herramientas disponibles en este SIG, con la finalidad de evaluar si las coordenadas son consistentes con la información geográfica del registro.

9. Aplicación de las herramientas disponibles en el SIG

   Es importante mencionar que se deberá tener claridad en el nivel administrativo al que se dirigirá el control de calidad. Esto puede ser definido con base en el tiempo disponible y la complejidad de la información geográfica contenida en la base de datos. Si el tiempo es limitado, se recomienda realizar el control de calidad en niveles administrativos superiores (*country*, *stateProvince*, *county*), ya que con las herramientas disponibles en los SIG es posible llevar a cabo esta tarea de forma relativamente sencilla.

En el siguiente capítulo, se explorará un campo de mayor dificultad debido a su naturaleza heterogénea en contenido y amplitud: el campo *locality*. Si se decide realizar el control de calidad a este nivel, se deberá contar con el tiempo suficiente para verificar que las coordenadas son consistentes con la descripción de localidad, ya que esta suele hacer referencia a distancias, direcciones, redes viales, rasgos geográficos, entre otros.

Si se decide llevar a cabo el control de calidad en niveles administrativos superiores (*country*, *stateProvince*, *county*), se puede recurrir a la herramienta de geoprocesamiento llamada "Intersección". Este algoritmo obtiene las secciones que concuerdan entre los objetos espaciales de las capas de entrada, que en este caso sería el archivo que contiene las coordenadas y los datos geográficos, y las de superposición, que serían las capas vectoriales de los catálogos (MG de INEGI para los datos nacionales y GADM para los internacionales). El resultado es una capa de intersección que contiene los atributos de los objetos que coinciden en ambas capas, de entrada y la superpuesta.

Esta capa resultante es la que se deberá utilizar para determinar si los datos son consistentes con lo que se indica en los campos *country*, *stateProvince* y *county*. Se sugiere que, previamente al control de calidad del dominio coordenadas, los datos del dominio geografía hayan sido procesados para que, al compararlos con los resultados de la intersección, se evite que sean inconsistentes por diferencias en el formato, errores ortográficos o tipográficos. Si los datos son iguales, entonces se considera que son consistentes. En caso contrario, se deberá determinar mediante el algoritmo Nearest Neighbour Join (NNJoin) la distancia a la cual se encuentran las coordenadas del país, estado o municipio indicado. Este algoritmo une las capas de vectores con base en relaciones del vecino más cercano, es decir, una entidad de la capa de entrada (coordenadas) se une a la entidad más cercana en la capa de unión (catálogos). En la tabla de atributos del catálogo se selecciona el nombre del nivel administrativo referido en la base de datos, se aplica el algoritmo NNJ y, como resultado, se obtiene una nueva capa vectorial que contiene los atributos de ambas capas y la distancia que existe entre la entidad de la capa de entrada y la entidad de la capa de unión. Si se encuentran a una distancia máxima de 2 km del nivel administrativo indicado para el registro (*country*, *stateProvince* o *county*) se consideran consistentes, de lo contrario se deberá evaluar si el dato indicado en la base de datos es incorrecto y deberá asignarse el valor correspondiente o, si la información geográfica es correcta, pero las coordenadas no lo son, se deberán calificar como inconsistentes.

Las coordenadas pueden discordar de los datos geográficos referidos (*stateProvince* o *county*), por dos motivos. El primero, es que el área del nivel administrativo ha sido modificada en el catálogo utilizado debido a una actualización, como la incorporación de un estado o un municipio), causando que las coordenadas se intersecan con un polígono cuyos datos no coinciden con lo indicado en la base de datos y se provoque esta disparidad. En este caso, las coordenadas son consistentes con la información geográfica a pesar de esta actualización, por lo que se deberá estandarizar el nombre del nivel administrativo a la

versión más actual del catálogo y calificarse como corresponde. El segundo motivo es que la distancia a la cual se encuentran las coordenadas del polígono del nivel administrativo referido es mayor a la distancia límite establecida de dos kilómetros, por lo que las coordenadas serían inconsistentes al no ser coherentes con los datos geográficos indicados.

Si se cuenta con el tiempo suficiente para realizar el control de calidad de las coordenadas a nivel de localidad, entonces el procedimiento continuaría de la siguiente manera:

Se deberán proyectar los catálogos, es decir, las capas de datos espaciales que contienen información a nivel de localidades. Inicialmente, se debe determinar el tipo de localidad descrita en *locality*, es decir, si se trata de un rasgo geográfico (una localidad, un cerro, una cueva, etcétera), si adicionalmente describe la distancia y la dirección (como 9 km N de Tlayacapan) o si se hace referencia a carreteras y kilometrajes (por ejemplo, Km 10 carretera Guadalajara-Chapala).

Rasgo geográfico

Para localidades que contienen un rasgo geográfico se utiliza la herramienta intersección, siguiendo el mismo procedimiento antes descrito para los niveles administrativos superiores, utilizando los catálogos correspondientes para *locality*. Aquellos registros que se intersecan son consistentes, de lo contrario se utiliza el algoritmo NNJ. Si la coordenada se encuentra a una distancia máxima de dos kilómetros del rasgo geográfico, entonces es consistente. Si supera dicha distancia, se deberá investigar si existe el lugar referido en otros catálogos o buscar bibliografía que indique la ubicación del mismo.

Distancia y dirección

Para localidades que tienen un rasgo geográfico, como punto de referencia a partir del cual se da una distancia y dirección, el procesamiento de las coordenadas es *manual*, ya que una vez que están proyectadas las coordenadas y los catálogos correspondientes, deberá utilizarse la herramienta "Medir línea" de QGIS para poder determinar si las coordenadas se ubican a la distancia y dirección indicadas en *locality*.

Carreteras y kilometrajes

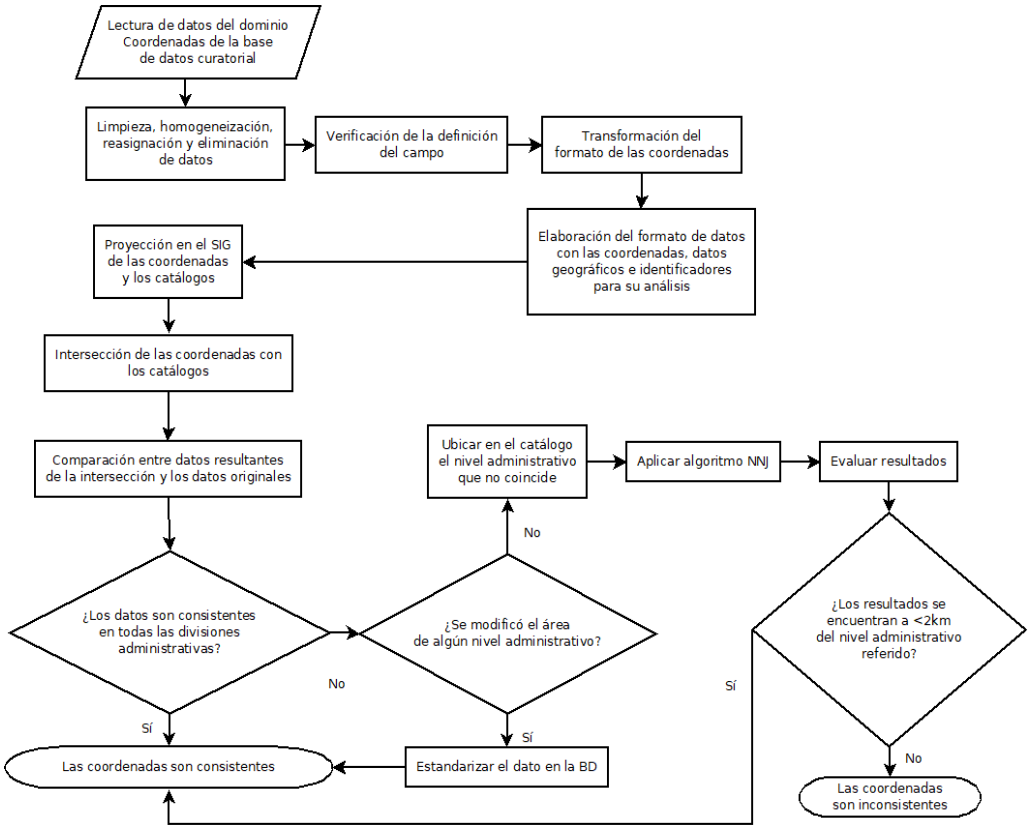
Así como en el caso anterior, este tipo de localidades deberá ser procesada de forma *manual* utilizando catálogos como la Red Nacional de Caminos incluida en los servicios WMS de INEGI, para los registros nacionales y GeoNames, Google Maps o bibliografía para registros internacionales. En este caso, se debe determinar si las coordenadas se encuentran en la carretera y kilómetro referido en *locality*, si es así, las coordenadas serán consistentes. Si las coordenadas no se ubican en la carretera y kilometraje referido o, en su defecto, dentro de los dos kilómetros de tolerancia, se calificarán como inconsistentes.

En todos los casos, si no se ubican las referencias de rasgos geográficos o carreteras, se deberá realizar el control de calidad de las coordenadas en los niveles administrativos

superiores a *locality* y se seguirá el procedimiento antes descrito, por lo que *locality*, *decimalLatitude* y *decimalLongitude* quedarán calificados como 0.6.

En la Figura 8, se muestra un diagrama de flujo que representa los pasos antes desarrollados.

Figura 7. Diagrama de flujo del proceso de control de calidad del dominio coordenadas



Nota. Los catálogos utilizados estarán definidos por el país de procedencia, si pertenecen a México, se empleará el Marco Geoestadístico del INEGI, utilizando los catálogos del nivel administrativo que corresponda. Si son internacionales, se utilizará de manera prioritaria el GADM en su versión más reciente. BD = Base de datos, NNJ = Nearest Neighbour Join.

Fuente: Elaboración propia.

Evaluación de la calidad de datos

El uso de calificaciones, es una herramienta esencial en la evaluación de la calidad del dominio coordenadas, pues hace posible determinar a través de valores numéricos el grado de confiabilidad que tienen los datos albergados en los campos de este dominio. Conocer estas calificaciones ayuda a los responsables de los datos a identificar áreas de oportunidad y a realizar las acciones pertinentes. Los puntos que se evalúan en el control de calidad de las coordenadas son los siguientes:

1. Limpieza de datos (eliminar partículas, dobles espacios, etc.) en los campos *verbatimLatitude* y *verbatimLongitude.*
2. Consistencia con los campos de geografía (si indica que pertenece a un determinado municipio, se deberá verificar que las coordenadas se ubiquen, en efecto, en dicho municipio).
3. Reasignación de datos que no correspondan a los campos correctos.
4. Conversión de las coordenadas *verbatim* a grados decimales.
5. Verificar que las coordenadas en *decimalLatitude* y *decimalLongitude* estén en grados decimales.

Cuadro 15. Calificaciones empleadas en la evaluación de la calidad de los datos del dominio coordenadas

	Calificación (qi)	*Situación*	*Ejemplos*
consistente	1	Consistente de origen: el campo original y procesado son iguales. Las coordenadas son consistentes con los demás datos geográficos disponibles en el registro (*country*, *stateProvince*, *county*, *locality*).	*decimalLatitude_o*: 19.31962 *decimalLongitude_o*: -99.19355 *decimalLatitude*: 19.31962 *decimalLongitude*: -99.19355
	0.9	Modificado por variantes de escritura (por ejemplo, digitación, ortografía, cambio de minúsculas o mayúsculas, completar y espacios). La identidad del valor se mantiene.	*verbatimLatitude_o*: 20° 23' 44.3'' W *verbatimLatitude*: 20° 23' 44.3'' N
	0.8	Valor asignado durante el proceso de control de calidad, es decir, no está presente en el campo original.	*decimalLatitude*: sin dato *decimalLongitude*: sin dato *decimalLatitude*: 19.31962 *decimalLongitude*: -99.19355
	0.7	Modificado por estandarización con respecto al catálogo.	Las coordenadas son transformadas de formato grados, minutos y segundos a grados decimales: *verbatimLatitude*: 19°1'36'' *verbatimLongitude*: -91°4'47'' *decimalLatitude*: 19.02667 *decimalLongitude*: -91.0797
	0.6	La consistencia de las coordenadas fue evaluada a nivel de geografía superior (*country*, *stateProvince*, *county*).	
inconsistente	0.4	Esta calificación suele asignarse cuando las coordenadas originales presentan algún tipo de problemática que no permite su conversión a grados decimales, y se requiere la resolución del proveedor de datos tras realizar la revisión en el control de calidad.	*verbatimLatitude_o*: 1883348.8218 UTM2N *verbatimLongitude_o*: 150483338.04 UTM1E
	0.1	Caso 1: Las coordenadas son inconsistentes con los datos geográficos descritos. Caso 2: Las coordenadas presentan datos solo en el campo de latitud o solo en el	Ejemplo caso 1: *country*: México Las coordenadas se ubican en Perú

		campo de longitud, impidiendo su procesamiento. Caso 3: El valor de las coordenadas excede el rango que, por definición, pueden abarcar.	Ejemplo caso 2: *verbatimLatitude*: 19° 06' 19'' N *verbatimLongitude*: sin dato Ejemplo caso 3: *verbatimLatitude*: 19°36' *verbatimLongitude*: -99°**82**' (el valor excede los 60 minutos)
nulo	0	El valor original se eliminó del campo para ser reasignado al campo correcto.	*verbatimLatitude_o*: Ciudad de México *verbatimLatitude*: sin dato
	Vacío (sin dato)	Cuando no existe valor para revisar.	*verbatimLatitude*: sin dato *verbatimLongitude*: sin dato *decimalLatitude*: sin dato *decimalLongitude*: sin dato

Nota. Las calificaciones se dividen en tres grupos: consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Perfil del personal

Para que el control de calidad del dominio coordenadas se lleve a cabo de manera efectiva, el personal a cargo deberá cumplir con una serie de conocimientos y habilidades, que abarcan desde una comprensión de conceptos biológicos básicos hasta la competencia en SIG y un dominio de los principios de geografía y cartografía. Asimismo, la capacidad para utilizar capas geoespaciales de fuentes confiables y aplicar habilidades de análisis de datos y resolución de problemas es primordial para la evaluación de la calidad de la información. La atención al detalle y las habilidades de comunicación efectiva son igualmente esenciales en este proceso colaborativo. Los puntos siguientes desarrollan las habilidades mínimas que debe cumplir el personal a cargo del control de calidad de este dominio.

1. Conocimientos en biología

   Es importante tener un entendimiento sólido (aunque general) de los conceptos biológicos para comprender la importancia de la precisión en los datos geográficos de las colecciones biológicas, así como el determinar si el dato es coherente con las características biológicas del ejemplar.

2. Competencia en SIG

   El personal encargado del control de calidad de las coordenadas debe tener experiencia y habilidades en el manejo de SIG. Esto implica la capacidad de utilizar el software SIG para visualizar, analizar y corregir los datos.

3. Conocimientos en geografía y cartografía

   Un buen entendimiento de los principios de la geografía y la cartografía es esencial para evaluar y corregir las coordenadas. Esto incluye conocimientos acerca de proyecciones cartográficas, sistemas de coordenadas, precisión cartográfica y técnicas de georreferenciación, así como la habilidad para emplear capas geoespaciales en diversos formatos.

4. Uso de capas geoespaciales

   El personal deberá de ser capaz de investigar, descargar y proyectar capas geoespaciales provenientes de fuentes fidedignas que puedan ser utilizadas para la evaluación de la consistencia de las coordenadas.

5. Habilidades en análisis de datos y resolución de problemas

   El personal debe ser capaz de analizar y evaluar datos geográficos de manera crítica, identificando posibles errores o inconsistencias. También deben tener habilidades en resolución de problemas para corregir y mejorar la calidad de los datos, ya sea mediante la verificación de fuentes adicionales, la depuración manual de registros o la utilización de técnicas avanzadas de procesamiento de datos geoespaciales.

6. Atención al detalle

   La precisión es fundamental en el control de calidad de los datos geográficos. Por lo tanto, el personal debe tener una gran atención al detalle y ser meticuloso al revisar y corregir las coordenadas. Si las coordenadas presentan errores, esto puede tener un impacto significativo en los análisis realizados por los usuarios de los datos.

7. Trabajo en equipo y comunicación

   El control de calidad de los datos geográficos a menudo implica colaborar con otros miembros del equipo, por lo tanto, es importante tener habilidades de trabajo en equipo y comunicación efectiva para colaborar en la resolución de problemas y asegurar la calidad de los datos.

Asignación de roles

El dominio de coordenadas necesita las labores de un coordinador y un analista especializado en metodología y estadística. En lo que respecta al control de calidad en este dominio, estas tareas son ejecutadas por un analista líder en datos geográficos y un analista de datos. A continuación, se detallan las actividades correspondientes:

1. Coordinador
2. Analista en metodología y estadística
3. Analista líder de datos geográficos

   Campos de los que es responsable: *verbatimLatitude*, *verbatimLongitude*, *decimalLatitude*, *decimalLongitude.*

   a. Resolución de casos especiales detectados por el analista de datos.
   b. Evaluación de la consistencia de las coordenadas con respecto al resto de los datos geográficos.
   c. Calificación de los campos *verbatimLatitude*, *verbatimLongitude*, *decimalLatitude*, *decimalLongitude.*

4. Analista de datos

   a. Asignación de datos que se encuentren en otros campos: *verbatimLatitude* y *verbatimLongitude.*

b. Reasignación de datos que no correspondan a los campos.
c. Limpieza de los datos de los campos *verbatimLatitude* y *verbatimLongitude.*
d. Si los datos de *verbatimLatitude* y *verbatimLongitude* se encuentran en grados decimales, asignación directa a *decimalLatitude* y *decimalLongitude.*
e. Evaluación sobre si las coordenadas se encuentran dentro de los límites válidos, así como su coherencia con su ubicación de forma general (por ejemplo, si el ejemplar indica que pertenece a México, las coordenadas deberían presentar una latitud con valores positivos y una longitud en valores negativos. Si se trata de coordenadas en formato sexagesimal, los minutos y segundos no son mayores a 60).
f. Transformación del formato de coordenadas de *verbatimLatitude* y *verbatimLongitude* (grados, minutos, segundos; grados, minutos) a grados decimales, en caso de ser necesario.
g. Asignación de coordenadas en el formato requerido de *verbatimLatitude* y *verbatimLongitude* a los campos *decimalLatitude* y *decimalLongitude*, en caso de que estos no cuenten con datos.
h. Detección de casos que requieren especial atención por parte del analista líder.

Localidad

El dominio localidad, en las colecciones biológicas, describe detalladamente la ubicación geográfica de las observaciones y especímenes recolectados. La consistencia de los datos en este dominio es importante, pues así se asegura la utilidad de la información y se garantiza que los datos geográficos sean confiables.

Este dominio posee información sumamente heterogénea, por lo que el *automatizar* su procesamiento resulta complejo. Los campos *locality* y *verbatimLocality* pueden contener una descripción sumamente detallada (como se muestra en el Ejemplo 1) que dificulta el procesamiento de cada elemento debido a la gran cantidad de información, hasta descripciones tan genéricas que resultan poco informativas (como se ilustra en el Ejemplo 2).

Ejemplo 1:
Cadena larga de localidad:

> "Village: Yoloxóchitl. Between: Yoloxóchitl and San Luis Acatlán. Vicinity: Lugar llamado El Ruidoso. entre Yoloxóchitl y San Luis Acatlán, como medio camino entre los dos pueblos, saliendo por una brecha hacia el norte de la carretera. Siguiendo como 200 mts despues de la primera colecta, también al lado derecho de la vereda. Ya otros 50 metros junto a un arroyo pequeño en área con el suelo muy aguado. Then following the brook another 150 or so meters. No continuing up a slope about 100 mts toward a path along a ridge, a path that that goes from Acatlán to Yoloxóchitl. And now. some 100 mts further. on the path from Yolo to Acatlán."

Ejemplo 2:
Cadena corta de localidad:

"Loma"

En el primer caso, se presenta una cantidad de información geográfica extensa y variada, que puede hacer complicada la tarea de filtrar y buscar datos específicos en la base de datos para su comparación con los catálogos utilizados, dificultando la estandarización de los datos. Esto puede provocar que se incluyan datos incorrectos o inadecuados, afectando la calidad de los datos. El procedimiento a seguir para el control de calidad de este dominio dependerá de la naturaleza de los datos contenidos en el mismo. Se sugiere que, si los datos de localidad son extensos, se limite el control de calidad a la limpieza, reasignación y eliminación de datos. En caso contrario, se puede realizar la validación del rasgo geográfico referido en la localidad.

Estándares de datos

En el dominio localidad, se siguen las recomendaciones establecidas por el estándar Darwin Core para los campos *verbatimLocality*, *locality* y *locationRemarks*. El campo *verbatimLocality* permite la preservación de los datos originales, aquí se conserva la descripción de la ubicación tal como se registró inicialmente en su forma sin procesar, manteniendo así un registro auténtico de los datos geográficos. Por otro lado, el campo *locality* alberga los datos estandarizados y estructurados de la localidad, garantizando su consistencia. Finalmente, el campo *locationRemarks* alberga información asociada a la localidad. Tal como en los otros dominios, el uso de los campos del estándar Darwin Core para la descripción de localidad, permite la interoperabilidad entre las bases de datos.

Base de datos

De acuerdo con los términos del estándar Darwin Core *verbatimLocality*, *locality* y *locationRemarks*, definidos en la sección anterior, se realiza el diseño de la base de datos en la que se llevará a cabo el control de calidad del dominio localidad. Se incluyen campos de control en los que se pueda indicar cuándo y quién realizó la última modificación en los datos, así como identificadores para cada uno de los registros:

Cuadro 16. Campos empleados en el control de calidad del Dominio Localidad

Tipo de campo	*Nombre del campo en Darwin Core*	*Estatus (obligatorio, opcional)*
Campos de registro	*verbatimLocality*	Opcional
	locality	Opcional
	locationRemarks	Opcional
Campos de control	*lastModified*	Obligatorio
	lastModifiedUser	Obligatorio
Identificadores	occurrenceID	Obligatorio
	uuid	Obligatorio

	datasetID	Obligatorio

Nota. En todos los "campos de registro" en la base de datos se incluyen campos con el mismo nombre que el campo de registro, más un sufijo para indicar los datos originales (sufijo "_o"), las calificaciones (con sufijo "_qi") y sus permisos o banderas (con sufijo "_pub").

Fuente: Elaboración propia con base en *Darwin Core Quick Reference Guide* (Grupo de Mantenimiento Darwin Core, 2021).

Catálogos

Los catálogos desempeñan funciones esenciales en el control de calidad del dominio localidad, que incluyen:

- Normalización de los nombres utilizados para describir los lugares, asegurando la consistencia en la descripción de la localidad.
- Evaluación de la pertenencia del rasgo geográfico al nivel administrativo indicado en los campos de país, estado o municipio.
- Evaluación de la consistencia de los datos geográficos.

El control de calidad de los datos de localidad implica la comparación de la información con catálogos provenientes de instituciones especializadas y fuentes confiables, lo que garantiza la precisión y la fiabilidad de los datos geográficos.

Cuadro 17. Catálogos utilizados para el dominio de localidad

Catálogo	*Temas utilizados*	*Formato*	*Observaciones*
MG INEGI	Polígonos de localidades urbanas y rurales	Shapefile	Registros nacionales
GADM	Niveles 3, 4 y 5	GeoPackage	Registros internacionales
WMS INEGI	Datos del relieve Información topográfica 1:50,000 Marco Geoestadístico Recursos naturales Red nacional de caminos Registro de nombres geográficos Territorio insular Zonas hidrogeológicas	Servicio de mapas Web (WMS)	Registros nacionales
Acervo histórico de localidades geoestadísticas del INEGI	NA	Valores separados por comas (CSV)	Registros nacionales
Conjunto de datos del Territorio Insular Mexicano. Escala 1:50,000.	NA	Shapefile	Registros nacionales

Versión 2.0			
GeoNames	NA	Base de datos geográfica (acceso a través de internet)	Registros nacionales e internacionales
Google Maps	NA	Servidor de aplicaciones de mapas web	Registros nacionales e internacionales

Fuente: Elaboración propia.

Además de los catálogos antes mencionados, un recurso muy utilizado para la validación de nombres de localidades es la bibliografía, ya que en ocasiones las localidades son tan puntuales que no forman parte de los catálogos disponibles en línea. Se sugiere la creación de un catálogo interno que contenga el nombre de la localidad, el país, estado, municipio y coordenadas (si es posible), así como la fuente de referencia.

Herramientas de datos

Se recomienda realizar una exploración inicial de la base de datos para evaluar la complejidad de las descripciones de la ubicación. Este análisis es crucial para determinar la estrategia a seguir en el control de calidad de este dominio, ya que las características de los datos y los desafíos asociados pueden variar significativamente de una colección a otra.

El control de calidad del dominio localidad se realiza de la siguiente manera:

1. Limpieza de los datos

   Se lleva a cabo una limpieza de datos de los campos *locality* y *verbatimLocality*, verificando que no presenten espacios, errores ortotipográficos, etc. Por ejemplo:

   viverosde Coyoacán
   Viveros de Coyoacán

2. Reasignación de datos

   Si existe información que no corresponde al campo *locality*, deberá ser reasignada al campo que le corresponda. Por ejemplo:

   Estación de Biología "Los Tuxtlas". Cerro El Vigía. En la copa de un árbol.
   Estación de Biología "Los Tuxtlas". Cerro El Vigía.

   a. *locationRemarks*

      Se han detectado registros en los que la localidad presenta datos que están asociados a la ubicación del evento, pero que no son informativos para la descripción de la localidad. Por ejemplo:

      Expuesta a sol directo, sin mantenimiento. Sta. Ma. Mazatla, Jilotzingo.
      Santa María Mazatla, Jilotzingo.

En este caso, la información deberá reasignarse a *locationRemarks*, conservando el valor indicado por el registro y acotando la información de *locality* a datos relevantes.

3. Eliminación de datos

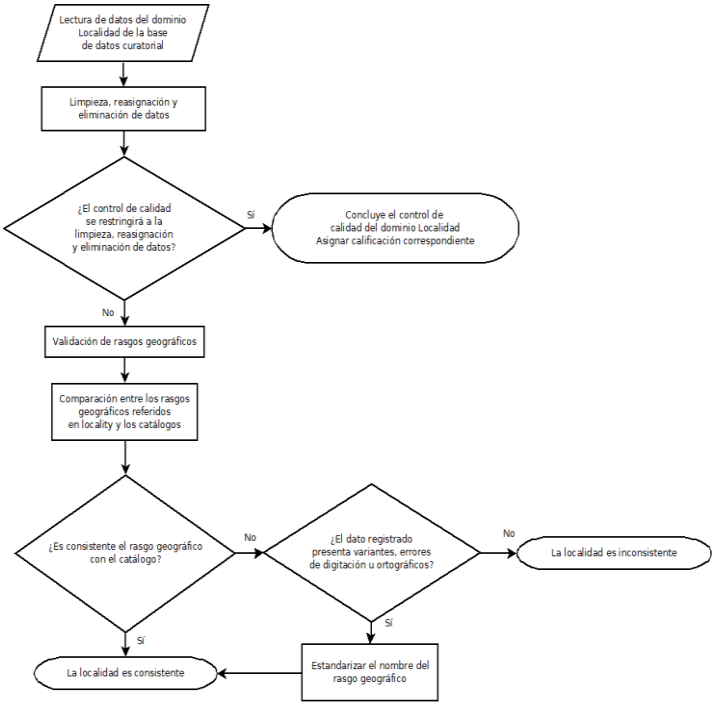
   Se eliminan los datos que no pueden ser reasignados y no son de carácter informativo para el registro en cuestión. Por ejemplo, la representación de datos nulos con valores (por ejemplo, 99999 9999 9999, @@@@, ND, SD), pues no son informativos y deberán ser eliminados.

4. Estandarización de rasgos geográficos

   Se identifican los rasgos geográficos mencionados en la localidad y compararlos con los catálogos para definir si estos se ubican en el país, estado y municipio indicados. Si persisten rasgos geográficos que no se pueden ubicar, el control de calidad se limitará a la limpieza, reasignación y eliminación de datos.

En la Figura 9, se muestra, a través de un diagrama de flujo, el procedimiento que se sigue en el control de calidad en el dominio localidad.

Figura 8. Diagrama de flujo del proceso de control de calidad del dominio localidad



Nota. Los catálogos utilizados se definirán según el país de procedencia. Si pertenecen a México, se empleará el Marco Geoestadístico del INEGI y se utilizarán los catálogos del nivel administrativo correspondiente. En caso de datos internacionales, se dará prioridad al uso de la versión más reciente del GADM.

Evaluación de la calidad de datos

Las calificaciones empleadas en el control de calidad de datos permiten conocer el estado de la colección en cuestión y también informan al responsable de la colección sobre las

modificaciones realizadas. En el dominio localidad las calificaciones pueden reflejar la complejidad de la descripción de localidad, ya que, por ejemplo, si el procesamiento de los campos se limita a una limpieza de datos (calificación igual a 0.6), se debe a que la información es sumamente heterogénea e impide su procesamiento de forma detallada. El control de calidad de la localidad se enfoca en la limpieza de los datos, la comparación del rasgo geográfico referido con los catálogos antes mencionados y la verificación de su correspondencia con los datos de país, estado o municipio indicados. Con base en este procesamiento son asignadas las calificaciones detalladas en la siguiente tabla:

Cuadro 18. Calificaciones empleadas en la evaluación de la calidad de los datos del dominio localidad

	Calificación (qi)	*Situación*	*Ejemplos*
consistente	1	Consistente de origen: el campo original y procesado son iguales.	*locality_o*: Chamela *locality*: Chamela
	0.9	Modificado por variantes de escritura (por ejemplo, digitación, ortografía, cambio de minúsculas o mayúsculas, completar y espacios). La identidad del valor se mantiene.	*locality_o*: **r**ío Copalita *locality*: **R**ío Copalita
	0.8	Valor asignado durante el proceso de control de calidad, es decir que no está presente en el campo original.	*locality_o*: sin dato *locality*: Ajijic
	0.7	Modificado por estandarización con respecto al catálogo.	*locality_o*: Zapotitlán de las Salinas *locality*: Zapotitlán Salinas
	0.6	El control de calidad se restringe a la limpieza de datos, se desconoce si los datos indicados en *locality* son consistentes con el resto de datos geográficos.	*locality_o*: 6 km. al NO deSan Pedro Nopala " Cerro Pericón". *locality*: 6 km al NO de San Pedro Nopala "Cerro Pericón".
inconsistente	0.4	Requiere la resolución del proveedor de datos tras realizar la revisión en el control de calidad.	La localidad presentaba un valor que presumiblemente correspondía a nombre científico e indicaba que era una localidad tipo: "**Anolis alvarezdeltoroi Type locality** (approximately), 19.5 km N, 8.1 km W Ocozocoautla"; sin embargo, el registro indica que su nombre científico es "Corytophanes hernandesii" y no se trata de ejemplares tipo
	0.1	No fue posible encontrar relación lógica ni con la definición del campo según el estándar ni se puede corroborar con ningún catálogo o referencia por lo que resulta inconsistente.	
nulo	0	El valor original se eliminó del campo para ser reasignado al campo correcto.	*locality_o*: selva *locality*: sin dato
	Vacío (sin dato)	Cuando no existe valor para revisar	*locality_o*: sin dato *locality*: sin dato

Nota. Las calificaciones se dividen en tres grupos: consistentes (mayores o igual a 0.6), inconsistentes (menores a 0.6) y nulo (sin dato).

Perfil del personal

El personal encargado del control de calidad del dominio localidad debe contar con diversas capacidades y competencias. Es importante que posea conocimientos en biología y sea capaz de detectar patrones de errores, así como manejar catálogos geográficos, habilidades de investigación y atención al detalle. En esta sección, se desarrollan las características del personal que realice el control de calidad del dominio localidad.

1. Conocimientos en biología

   El personal encargado del control de calidad de las localidades debe ser capaz de evaluar si la información de la localidad es coherente con la distribución general conocida del organismo y detectar posibles errores o discrepancias.

2. Capacidad de detección de patrones

   Dada la heterogeneidad del campo de *locality*, es importante que el personal sea capaz de identificar patrones de errores que puedan ser corregidos en bloques que abarquen la mayor cantidad de registros como sea posible, agilizando así el control de calidad de este complejo campo.

3. Conocimientos en geografía y cartografía

   Un buen entendimiento de los principios de la geografía y la cartografía es esencial para evaluar y corregir las localidades en las bases de datos. Esto incluye conocimientos sobre sistemas de coordenadas, nombres de lugares y técnicas de georreferenciación.

4. Habilidades en investigación y verificación de datos

   El personal debe ser capaz de investigar y verificar la información de la localidad utilizando diversas fuentes, como literatura científica, mapas históricos, registros climáticos y otros recursos geográficos. Además, deben tener habilidades para identificar y corregir posibles errores o inconsistencias.

5. Atención al detalle

   La precisión y la atención al detalle son fundamentales en el control de calidad de las localidades. El personal debe ser meticuloso al revisar y corregir la información, asegurándose de que la información sea coherente con los demás campos que contienen los datos geográficos, así como con las características biológicas del ejemplar.

6. Competencia en SIG

   El conocimiento y la experiencia en el uso de software SIG, permiten visualizar y analizar las localidades, así como la proyección de los catálogos utilizados.

7. Trabajo en equipo y comunicación

   El control de calidad de las localidades a menudo implica colaborar con otros profesionales, por lo tanto, es importante tener habilidades de trabajo en equipo y comunicación efectiva para colaborar en la resolución de problemas y garantizar la calidad de los datos.

Asignación de roles

Las funciones de un coordinador y un analista especializado en metodología y estadística son necesarias en el dominio localidad. Las actividades del control de calidad en este dominio las realizan un analista líder en datos geográficos y un analista de datos. A continuación, se describen las tareas correspondientes:

1. Coordinador
2. Analista en metodología y estadística
3. Analista líder en datos geográficos
    a. Análisis de los datos para definir el tipo de localidad.
    b. Comparación con catálogos especializados para determinar la coherencia de los datos.
    c. Estandarización de los nombres de las localidades.
    d. Evaluación de la consistencia de la localidad con respecto al resto de los datos geográficos.
    e. Calificación de los campos *locality* y *verbatimLocality*.
4. Analista de datos
5. Limpieza de los datos *locality* y *verbatimLocality*
6. Reasignación de datos que no correspondan a los campos correctos

BIBLIOGRAFÍA

Arlé, E., Zizka, A., Keil, P., Winter, M., Essl, F., Knight, T., Patrick, W., Jiménez-Muñoz, M. & Meyer, C. (2021). bRacatus: A method to estimate the accuracy and biogeographical status of georeferenced biological data. *Methods in Ecology and Evolution, 12*(9), 1609-1619.

Bánki, O., Roskov, Y., Döring, M., Ower, G., Hernández Robles, D. R., Plata Corredor, C. A., Stjernegaard Jeppesen, T., …, Şentürk, O. (2023). Catalogue of Life Checklist (Annual Checklist 2023). *Catalogue of Life*. https://doi.org/10.48580/dfsr.

Bell, F. W., Kershaw, M., Aubin, I., Thiffault, N., Dacosta, J., & Wiensczyk, A. (2011). Ecology and traits of plant species that compete with boreal and temperate forest conifers: An overview of available information and its use in forest management in Canada. *The Forestry Chronicle, 87*(2), 161-174.

Bertone, R. & Thomas, P. (2011). Introducción a las bases de datos. Fundamentos y Diseño. *Pearson HispanoAmerica Contenido, Buenos Aires*.

Canadensys. (s.f.). Coordinate conversion. *Canadensys*. Recuperado el 16 de junio de 2023 de https://data.canadensys.net/tools/coordinates.

Castillo, M., Michán, L. y Martínez, A. L. (2014). La biocuración en biodiversidad: proceso, aciertos, errores, soluciones y perspectivas. *Acta Botánica Mexicana, 108*, 81-103. https://www.scielo.org.mx/scielo.php?pid=S0187-71512014000300006&script=sci_abstract.

CCUD, UNAM. (2017). *Manual de Datos Abiertos*. Ciudad Universitaria, CDMX. Universidad Nacional Autónoma de México, pp. 80-81.

Chapman, A.D. & Wieczorek, J.R. (2020). Georeferencing Best Practices. *GBIF Secretariat*. https://docs.gbif.org/georeferencing-best-practices/1.0/en/.

Clasificación de Especies Colombia, Ministerio del Medio Ambiente. https://clasificacionespecies.mma.gob.cl/.

COAR, (2018). Confederation of Open Access Repositories, https://www.coar-repositories.org/files/coar-cv-infog-f_27051415-2.pdf.

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) (2019). *Datos primarios de ejemplares del Sistema Nacional sobre Biodiversidad (SNIB) – características y reglas –*. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. Ciudad de México. www.snib.mx/ejemplares/docs/CONABIO-SNIB-ProtocoloCalidadI.pdf.

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) (comp.) (2023). Catálogo de autoridades taxonómicas de especies de flora y fauna con distribución en México. Base de datos. SNIB-CONABIO, México. https://www.snib.mx/taxonomia/descarga/.

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. (2017). Regiones Marinas [Mapa]. 1:250000. *CONABIO*. http://geoportal.conabio.gob.mx/metadatos/doc/html/regionmarinamx.html.

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. (2021). *Guía de georreferenciación de localidades de Colecciones Biológicas* [PDF].

https://www.biodiversidad.gob.mx/media/1/conabio/documentos/proyectos/guia_georreferenciacion_2021_WEBc.pdf.

Confederation of Open Access Repositories. (2015). COAR Current State and Future Trends. Version 1 [PDF]. https://www.coar-repositories.org/files/coar-cv-infog-f_27051415-2.pdf.

Dalcin, E. C. (2005). Data quality concepts and techniques applied to taxonomic databases (Doctoral dissertation, University of Southampton). https://doi.org/10.13140/2.1.4440.2562.

Darwin Core Maintenance Group (2021). Darwin Core Quick Reference Guide. *Biodiversity Information Standards (TDWG)*. https://dwc.tdwg.org/terms/.

Darwin Core Task Group (2009). Darwin Core. *Biodiversity Information Standards (TDWG)*. http://www.tdwg.org/standards/450.

Delpeuch, A. (29 de diciembre de 2022). OpenRefine user manual. *OpenRefine*. https://openrefine.org/docs.

Dirección General de Repositorios Universitarios (DGRU) (2019). Estándar para Datos de Biodiversidad Darwin Core (DWC) (Versión 1). México: *Dirección General de Repositorios Universitarios. SDI-UNAM*. https://dgru.unam.mx/wp-content/uploads/2019/10/D.ST_.DGRU_CDI_007_2015_E_Datos_Biodiversidad_Darwin_Core.pdf.

Dirección General de Repositorios Universitarios (DGRU) (2022). Procedimiento de operación para el control de calidad de datos de colecciones a integrarse al Portal de Datos Abiertos, UNAM Colecciones Universitarias (PO.DGRU/CDI/002/2016/G en su versión vigente). México: Dirección General de Repositorios Universitarios. SDI-UNAM.

Esri. (s.f.). ¿Qué son las tablas y la información de atributos? https://desktop.arcgis.com/es/arcmap/latest/manage-data/tables/what-are-tables-and-attribute-information.htm.

Esri. (s.f.). Conceptos básicos de SIG (Sistemas de Información Geográfica). https://www.esri.com/es-es/what-is-gis/overview.

Food and Agriculture Organization of the United Nations. (s.f.). Buscar datos en formato shapefile. https://data.apps.fao.org/map/catalog/static/search?format=shapefile.

GBIF (s.f.). ¿Qué es Darwin Core y por qué es importante? https://www.gbif.org/es/darwin-core.

GBIF.org (2023). Página de Inicio de GBIF. https://www.gbif.org/es/.

Geoinnova. (11 de agosto de 2021). ¿Qué es un SIG, GIS o Sistema de Información Geográfica? *Geoinnova*. https://geoinnova.org/blog-territorio/que-es-un-sig-gis-o-sistema-de-informacion-geografica/.

GeoNames. (s.f.). Acerca de GeoNames. https://www.geonames.org/about.html.

GeoNames. (s.f.). GeoNames.]https://www.geonames.org/.

Global Administrative Areas (2023). *University of California, Berkeley*. [digital geospatial data]. http://www.gadm.org.

Godínez, J. L. (2008). Colectores de algas de México (1787-1954). *Acta Botánica Mexicana, 85*, 75-97. https://www.scielo.org.mx/pdf/abm/n85/n85a6.pdf.

Grupo de Mantenimiento Darwin Core. (2021). *Guía de referencia rápida de Darwin Core*. Normas de Información sobre Biodiversidad *(TDWG)*. https://dwc.tdwg.org/terms/.

Guiry, M.D. & Guiry, G.M. (2023). *AlgaeBase*. World-wide electronic publication: National University of Ireland, Galway. https://www.algaebase.org/.

IBM s.f., ¿Qué es PostgreSQL? https://www.ibm.com/mx-es/topics/postgresql.

INEGI. (s.f.). Marco Geoestadístico. https://www.inegi.org.mx/temas/mg/.

INEGI. (s.f.). Servicio de Información Geográfica. https://www.inegi.org.mx/servicios/wsinfogeo/default.html.

Instituto Nacional de Estadística y Geografía. (s.f.). Servicio de Mapas Web. https://www.inegi.org.mx/servicios/wsinfogeo/default.html.

International Commission on Zoological Nomenclature (ICZN) (1999). International Code of Zoological Nomenclature. Fourth edition. *The International Trust for Zoological Nomenclature*. https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/.

International Organization for Standardization [ISO] (2019). Date and time. Representations for information interchange – Part 1: Basic rules (ISO 8601-1:2019). https://www.iso.org/iso-8601-date-and-time-format.html.

International Organization for Standardization. (s.f.). Glossary for ISO 3166. https://www.iso.org/glossary-for-iso-3166.html.

International Organization for Standardization. (s.f.). ISO 3166 Country Codes. https://www.iso.org/iso-3166-country-codes.html.

IPNI (2023). The International Plant Names Index. *The Royal Botanic Gardens, Kew, Harvard University Herbaria & Libraries and Australian National Herbarium*. http://www.ipni.org/.

ITHAKA (2000-2023). Plant Collectors. *JSTOR Global Plants*. https://plants.jstor.org/collection/PPERSBM.

Lorenzo, C, Espinoza, E., Briones y Cervantes, F. A. (Eds.) (2006). Colecciones mastozoológicas de México. *Instituto de Biología, Universidad Nacional Autónoma de México y Asociación de Mastozoología, A.C.* http://mamiferosmexico.org/books/Colecciones_mastozoologicas.pdf.

Muñoz, D. E., Martínez, D. E., Márquez, R., Ramírez, E., García E., Juárez, F. A., Villanueva F. y Osorio, A. (Eds.) (2016). Metodología de limpieza de datos con la herramienta de OpenRefine. Centro Universitario de Liderazgo y Tecnología Avanzada. http://www.visualix.mx/files/5eb360388714341c277936ecb036e370.pdf.

Murguía-Romero, M., Ortíz-Bermudez, E., Serrano-Estrada, B., y Villaseñor-Ríos, J. L. (2022). Main collectors of Mexico's vascular plants: a catalogue built from online databases. *Revista Mexicana de Biodiversidad, 93*, e934044. https://doi.org/10.22201/ib.20078706e.2022.93.4044.

National Museum of Natural History & Smithsonian Institution (2023). Integrated Taxonomic Information System (ITIS). https://doi.org/10.5066/F7KH0KBK.

Ortega, S. y Guevara, A. (2017, 3-5 de julio). Darwin Core: Estándar para la Gestión de Datos Biológicos Primarios en la UTN. [primer]. Encuentro Latinoamericano de eCiencia, San José, Costa Rica.

https://documentas.redclara.net/bitstream/10786/1296/1/Darwin%20Core%20Est%C3%A1ndar%20para%20la%20Gesti%C3%B3n%20de%20Datos%20Biol%C3%B3gicos.pdf.

PostgreSQL Tutorial Website (2022). PostgreSQL EXTRACT Function. *PostgreSQL Tutorial*. https://www.postgresqltutorial.com/postgresql-date-functions/postgresql-extract/.

QGIS project. (2022). Una introducción fácil a GIS: 3. Datos Vectoriales. *QGIS Documentation*. https://docs.qgis.org/3.28/es/docs/gentle_gis_introduction/vector_data.html.

QGIS. (s.f.). Introducción suave a los datos vectoriales. https://docs.qgis.org/3.28/es/docs/gentle_gis_introduction/vector_data.html#:~:text=Datos%20vectoriales%20se%20utilizan%20para,de%20atributos%20que%20lo%20describen.

QGIS. (s.f.). QGIS - El SIG Líder de Código Abierto para Escritorio. *QGIS*. https://qgis.org/es/site/about/index.html.

Radziwill, Nicole M. Foundations for Quality Management, *Quality Management Journal, 2006*, 7, doi: 10.1080/10686967.2006.11918546.

Rzedowski, J., Calderón de Rzedowski, G., & Butanda, A. (2009). Los principales colectores de plantas activos en México entre 1700 y 1930. Instituto de Ecología, A.C. y Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. https://bioteca.biodiversidad.gob.mx/janium/Documentos/6333.pdf.

The Plant List. (2013). The Plant List. A working list of all plant v.1.1. http://www.theplantlist.org/.

The President and Fellows of Harvard College. (2013). Index of Botanists. Harvard University Herbaria & Libraries. https://kiki.huh.harvard.edu/databases/botanist_index.html.

Torres-Mejía, M., Beltrán, N., & Llano, S. (2016). Listas de especies: Lineamientos conceptuales y metodológicos para su consolidación en Colombia. Sistema de Información sobre Biodiversidad de Colombia, Bogotá D.C., Colombia. http://repository.humboldt.org.co/bitstream/handle/20.500.11761/9844/16-181-final.pdf?sequence=6&isAllowed=y.

Tropicos. (2023a). Tropicos v3.4.1. Missouri: Tropicos.org. Missouri Botanical Garden. https://tropicos.org/home.

Tropicos. (2023b). Tropicos v3.4.1. Missouri: Tropicos.org. Missouri Botanical Garden. Person Search https://tropicos.org/person/Search.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J., & Smith, G. F. (Eds.) (2018). International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile 159. Glashütten: Koeltz Botanical Books. https://doi.org/10.12705/Code.2018.

Universidad Nacional Autónoma de México. (2015). Datos de biodiversidad - Darwin Core [PDF]. https://dgru.unam.mx/wp-

content/uploads/2019/10/D.ST_.DGRU_CDI_007_2015_E_Datos_Biodiversidad_Darwin_Core.pdf.

Universidad Nacional Autónoma de México. (s.f.). Revista UNAM. Vol. 17, Num. 12, Artículo 87. https://www.revista.unam.mx/vol.17/num12/art87/.

Veiga, A. K., Saraiva, A. M., & Cartolano, E. A. (2014). Data quality control in biodiversity informatics: the case of species occurrence data. IEEE Latin America Transactions, 12(4), 683-693. https://doi.org/10.1109/TLA.2014.6868870.

WFO. (2023). The World Flora Online v.2023.01. http://www.worldfloraonline.org.

Wieczorek, J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni, R., Robertson, T., & Vieglais D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE, 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715.

WoRMS Editorial Board. (2023). World Register of Marine Species. https://www.marinespecies.org at VLIZ https://doi.org/10.14284/170.

GLOSARIO

Análisis exploratorio de la base de datos. Serie de revisiones y análisis que se realizan para examinar el contenido de las bases de datos a fin de que se pueda corroborar si los datos son correctos, anómalos o si contienen algún tipo de problemas, como valores inexactos o incongruentes.

Analista de datos. Analista encargado de la revisión, análisis y validación de conjuntos temáticos de información proveniente de las bases de datos originales. Por ejemplo, datos taxonómicos, datos de colecta o datos geográficos.

Analista en metodología y estadística. Analista responsable de integrar las bases originales proporcionadas por los proveedores de datos. Esto implica la creación de la base de datos curatorial y generación de tablas de trabajo especializadas para cada analista de datos, adaptadas al tipo de información que se revisará.

Analista líder. Analista que revisa los datos analizados y validados por el analista de datos y que está en coordinación con el analista en metodología y estadística para la migración de los datos validados a la base de datos curatorial.

Atomizar. Procedimiento en el que se separan los elementos de una cadena de texto presente en un campo en particular por medio de elementos específicos identificables dentro de la cadena de texto. Por ejemplo, espacios, signos de puntuación, palabras u otros símbolos específicos.

Atributos. Información asociada a la capa de datos espaciales. Por ejemplo, nombres, valores, códigos, fechas.

Base de datos curatorial. Corresponde con la base de datos que almacena todos los registros y campos de un conjunto de datos como una colección que se ingresa a la Plataforma Informática Aurora ® y que únicamente puede consultarse por medio de sistemas que requieren nombre de usuario y contraseña. La Plataforma informática Aurora ® es un sistema interoperativo desarrollado por la DGRU, que está compuesta por bases de datos estandarizadas y servicios web, que conjuntan datos de colecciones, capas geoespaciales y objetos digitales.

Base de Datos de Áreas Administrativas Globales (GADM). Colección de datos vectoriales que cubre unidades administrativas en todo el mundo, incluyendo los niveles políticos de cada país y sus subdivisiones.

Capa de datos espaciales. Representación digital de objetos geográficos del mundo real. Almacena información geográfica.

Catálogos especializados. Son conjuntos de datos especializados que contienen información que ha sido confirmada o respaldada por expertos en campos particulares y se emplean como fuentes confiables en la validación de los datos.

Colecciones Biológicas. Conjunto de ejemplares provenientes de las colectas biológicas. Se conforman por ejemplares, sus partes o derivados, preservados por distintos medios (físicos o químicos) y son organizados para integrar acervos que contienen información valiosa que coadyuva en el conocimiento de la biodiversidad.

Concatenar. Procedimiento en el cual la información contenida en dos o más campos de un registro se combina para generar información conjunta que puede ser asignada a otro campo contenedor. Por ejemplo, la clasificación superior en taxonomía o geografía o los elementos de una fecha.

Consulta de actualización. Serie de comandos o fórmulas que se emplean para asignar la información a cada campo después de haber realizado la comparación con un catálogo, por lo que la información que se asigna desde el catálogo corresponde al valor validado por ese campo en un registro.

Consulta de atomización. Serie de comandos o fórmulas empleados para separar los elementos de un campo en dos o más según el tipo de valor del que se trate. Por ejemplo, la atomización de un nombre científico de una especie en sus componentes como género, especie y autor.

Consulta de comparación. Serie de comandos o fórmulas que se emplean para comparar la información entre los campos de revisión de los registros y los catálogos. Este proceso determina el grado de consistencia de la información en los registros.

Consulta de concatenación. Serie de comandos o fórmulas que se emplean para realizar la unión de uno o más campos de un registro en uno solo.

Consulta de selección con parámetros. Serie de comandos o fórmulas que se emplean para identificar patrones dentro de la información de los campos de los registros de la base de datos que pueden incluir la identificación de símbolos, letras, números, palabras o frases que permiten agrupar registros similares entre sí debido a que comparten información similar.

Control de calidad en datos. Los procedimientos, actividades, recursos y evaluaciones utilizados para identificar posibles errores en los registros y campos, con el propósito de asegurar que una base de datos o conjunto de datos cumpla con las especificaciones definidas y se mantenga consistente a lo largo del tiempo.

Dato. Información que contiene la celda de un campo para un registro dado. Puede ser desde un número, letra, símbolos, palabras o descripciones más amplias.

Datos vectoriales. Tipo de representación geoespacial de información que utiliza geometría vectorial para describir objetos y fenómenos en el mundo real. Los datos vectoriales se representan mediante geometrías de tipo punto, línea o polígonos (por ejemplo, una carretera puede representarse mediante una línea, un lago por un polígono, un árbol mediante un punto).

Definición del campo. Corresponde a cómo se define una categoría de información que se encuentra almacenada en una tabla que pertenece a una base de datos. La definición del campo indica el tipo de dato que puede contener y cómo debe ser ingresado y expresado para tener el significado correcto.

Denominador de campo. Nombre que permite identificar un tipo de dato de otro dentro de una etiqueta de un ejemplar. Los tipos de denominadores varían según el tipo de dato que designen y algunas veces pueden presentarse abreviados.

Diagnóstico de control de calidad. Es un tipo de informe en el que se describe el estado de la base de datos original proporcionado por el proveedor de datos. Puede contener pero no estar limitado a una serie de listados de valores (tablas de datos), estadísticas (tablas de datos y/o gráficas), el análisis de correspondencia con catálogos (calificación de la calidad de los datos), adecuaciones realizadas sobre los datos (necesarias para la garantizar la interoperabilidad de los datos), así como tablas en las que se registra la conformidad de las propuestas presentadas al proveedor de los datos así como sus observaciones sobre el proceso del diagnóstico.

Ejemplar. Un espécimen es el material biológico asociado a un evento de colecta y puede consistir en uno o más organismos biológicos, montados o fijados en un medio físico para su preservación. Por ejemplo, una planta herborizada en una cartulina, o un animal en un frasco con un líquido que sirve de fijador de los tejidos.

Error de asignación. Ocurre cuando se captura información correspondiente a un campo particular en otro que no le corresponde.

Error de asignación total. Se presenta cuando se captura información que no está presente en la descripción propia del objeto

Error de captura. Fallas que son generadas en la incorporación de información a la base de datos, que puede incluir desde imprecisión de los datos, inconsistencias ortográficas o tipográficas, abreviaturas, omisiones, asignaciones, así como una mala atomización de la información, entre otras.

Error de digitación. Se presenta cuando en la captura de datos se han agregado, cambiado u omitido caracteres tales como letras, números, símbolos, etc. Los que errores de tipo ortográficos están incluidos en esta categoría.

Error de omisión. Se genera cuando cualquier elemento presente en la descripción del objeto no es capturado, tales como palabras, frases, enunciados, párrafos, o datos completos, etc.

Error de origen. Inconsistencias en la información que derivan directamente del objeto que se describe.

Error de proceso. Deficiencias en el procesamiento de la información generadas en alguno de los pasos durante los procesos tanto manuales como automáticos que son posteriores a la captura de información.

Estándar Darwin Core. Es un cuerpo de normas que incluye un glosario de términos (en otros contextos éstos se pueden llamar propiedades, elementos, campos, columnas, atributos o conceptos) destinado a facilitar el intercambio de información sobre la diversidad biológica, proporcionando definiciones de referencia, ejemplos y comentarios. Fue diseñado para facilitar la recuperación e integración de datos primarios que documenten la presencia u ocurrencia de especímenes biológicos modernos en el espacio y tiempo, así como las diferentes colecciones (físicas o digitales).

Estándar de datos. Es un modelo de datos que funciona como un punto de referencia para los campos que deben integrarse en una tabla o base de datos, en el que se determina tipo de datos que deben capturarse para conjuntos de datos que son de la misma naturaleza, por cual permite que puedan integrarse y ser consultadas como si se tratara de una sola base de datos fuente.

Estandarización. Conjunto de procesos por los cuales los registros de la misma naturaleza se convierten a un formato homogéneo, tomando como base una fuente documental.

Expresión regular. Cadena de caracteres que se usa para describir e identificar patrones comunes de coincidencia entre cadenas de texto. Por ejemplo: todos los valores de un campo que comienzan con una letra mayúscula o todos aquellos valores que incluyen algún número.

Geometría. Forma que representa un objeto espacial del mundo real. Se conforma por la interconexión de uno o más vértices.

GeoNames. Base de datos geográfica disponible de forma gratuita para su descarga. Integra datos geográficos, como nombres de lugares en varios idiomas, elevación, población y otros, de diversas fuentes. Todas las coordenadas de latitud y longitud están en el sistema WGS84 (Sistema Geodésico Mundial 1984).

Homogeneización. Proceso por el cual los registros de la misma naturaleza y presentes en varias bases de datos se convierten a un formato homogéneo, cuando no se tienen fuentes documentales específicas como catálogos

Integración de datos. Proceso por el cual se unifican los registros de una misma fuente en una sola base de datos para poder ser revisados y analizados.

Limpieza de datos. Conjunto de procesos y revisiones encaminadas a la detección y corrección de los datos "sucios", en los que se identifican causas y se categorizan en función de su naturaleza tales como: datos duplicados, faltantes o incompletos, inconsistentes o erróneos, así como datos que son inutilizables debido al estado de su calidad.

Marco Geoestadístico (MG). Sistema nacional diseñado por el Instituto Nacional de Estadística y Geografía (INEGI) que muestra las divisiones geoestadísticas del territorio continental e insular en diversos niveles de detalle, permitiendo ubicar de manera geográfica la información estadística proveniente de censos, encuestas institucionales y datos de las Unidades del Estado.

Migración de datos. Proceso de transferencia de la información de una base de datos de origen a otra de destino que tiene un diseño y estructura diferente. Por lo general se aplica para la base original proporcionada por el proveedor de datos.

Nivel administrativo. Jerarquía o estructura de división territorial utilizada para la gestión y organización de áreas geográficas a nivel mundial y dentro de un país o región (p. ej. en el catálogo GADM el nivel administrativo 0 es el nivel más alto en jerarquía y corresponde a país; el nivel 1 corresponde a divisiones estatales, provincias y equivalentes; y el nivel 2 a divisiones más pequeñas como municipios).

Normalización. Establecimiento de criterios para la reducción de variantes en la información de los campos de una base de datos con el fin de minimizar la redundancia de datos, lo que permite una mejor gestión posterior de los mismos. Puede incluirse la estandarización y la homogeneización dependiendo del tipo de información contenida en los campos y las normas o catálogos establecidos según su propia definición para su revisión.

Rasgo geográfico. Localidad principal que puede estar indicada por el nombre de una población, formaciones geológicas (por ejemplo, volcanes, cerros, montaña, cañón, cueva, islas, archipiélagos), carretera, instituciones, o cualquier otra ubicación referida en la localidad.

Registro (de ejemplar). Conjunto de datos o celdas relacionados entre sí, que constituyen una unidad o fila de información en una base de datos.

Servicio de Mapas Web (WMS). Un WMS es un estándar para publicar cartografía en línea según las especificaciones del Open Geospatial Consortium (OGC). Se utiliza para consultar información cartográfica a través de Sistemas de Información Geográfica (SIG) en computadoras de escritorio o para crear aplicaciones web híbridas (Mashups).

Sistema de Información Geográfica. Herramienta que se encarga de gestionar, analizar y visualizar datos en forma de mapas, combinando información sobre ubicación con detalles descriptivos de los elementos en esa ubicación. Vincula datos geoespaciales con información contextual.

Tabla de trabajo. Conjunto de bases de datos en las que se preserva el registro detallado de cada una de los procedimientos y revisiones llevadas a cabo en la gestión y validación de las bases de datos; las cuales incluyen todos los cambios realizados durante el proceso de control de calidad, lo cual facilita la recuperación de los valores originales de un registro si el proveedor de datos desea compararlo con su valor validado.

Validación. Proceso que permite determinar tanto la integridad, precisión y exactitud de la información contenida en los registros de la base datos, así como su congruencia, coherencia y pertinencia; lo que determinará el estado de potencial de uso posterior. La información se valida con el uso de catálogos de autoridad según el tipo de dato, e igualmente se emplean los criterios descritos por el estándar de datos y acuerdos establecidos con el proveedor de datos.

Valor. Un valor es la interpretación para una secuencia de información contenida en un dato, es decir, el significado que el dato tiene al relacionarlo con el nombre de la celda y como está relacionado con la naturaleza del registro al que pertenece.

Vocabularios controlados. Es un listado de palabras y términos que se emplean para delimitar el tipo de información que se introduce en una categoría de información, que está definido por la naturaleza intrínseca de la información que contiene dicha categoría.

ANEXO I

Índice temático de los servicios WMS del INEGI utilizados en el control de calidad de los datos geográficos de las colecciones biológicas (Acervo de información Geográfica INEGI (Mapa Digital de México v6.1)).

Nombre	*Título*
Fenómenos Geológicos	Volcán
Geodesia	Red geodésica nacional pasiva
Información topográfica 1:50,000	Acueducto Aeropuertos Canal Cuerpos de agua Presas Textos de cuerpos de agua Vías férreas
Marco Geoestadístico	Áreas geoestadísticas estatales Calles (nombres) Límite municipal Localidad rural Localidad urbana
Recursos naturales	Arrecifes
Red nacional de caminos	Caminos revestidos Caminos terracería Carretera N/A Carreteras y calles Casetas de peaje Postes de referencia
Registro de nombres geográficos	Áreas naturales y culturales Nombres de localidades Obras de infraestructura Rasgos hidrográficos Rasgos orográficos
Territorio insular	Arrecifes Oceánicos costeros Terrestres (islas en cuerpos de agua continentales)-2013
Zonas hidrogeológicas	Manantiales