# OPENREFINE

## GUÍA DE VALIDACIÓN Y LIMPIEZA DE DATOS SOBRE BIODIVERSIDAD

**JULIO- 2019**

*Versión - 1.0*

Sistema de Información sobre Biodiversidad de Colombia

**Cítese como:** SiB Colombia (2019). *OpenRefine - Guía de validación y limpieza de datos sobre biodiversidad*. Sistema de Información sobre Biodiversidad de Colombia, Bogotá D.C., Colombia, 12 pp. Disponible en: http://hdl.handle.net/20.500.11761/35350



**Control del documento:**

| Versión | Descripción | Fecha publicación | Autor(es) |
|---|---|---|---|
| 1.0 | Creación del documento | 2019.07.02 | Ricardo Ortiz, Camila Plata, Leonardo Buitrago |

Sistema de Información sobre Biodiversidad de Colombia

# Acerca del SiB Colombia

**El SiB Colombia es la red nacional de datos abiertos sobre biodiversidad.** Esta iniciativa de país nace con el Decreto 1603 de 1994 como parte del proceso de creación del Sistema Nacional Ambiental (SINA), establecido en la Ley 99 de 1993, y es el nodo oficial del país en la Infraestructura Mundial de Información en Biodiversidad (GBIF). Su principal propósito es brindar acceso abierto a información sobre la diversidad biológica del país para la construcción de una sociedad sostenible. Además, facilita la publicación en línea de datos e información sobre biodiversidad, y promueve su uso por parte de una amplia variedad de audiencias, apoyando de forma oportuna y eficiente la gestión integral de la biodiversidad.

El SiB Colombia es una realidad gracias a la participación de cientos de organizaciones y personas que comparten datos e información bajo los principios de libre acceso, transparencia, cooperación, reconocimiento y responsabilidad compartida.

Lo coordina el Instituto Humboldt y es liderado por un Comité Directivo (CD-SiB), conformado por el Ministerio de Ambiente y Desarrollo Sostenible, los 5 institutos de investigación del SINA (Ideam, Invemar, IIAP, Sinchi e Instituto Humboldt), la Universidad Nacional de Colombia y Parques Nacionales Naturales de Colombia. El CD-SiB se apoya en un Comité Técnico (CT-SiB), grupos de trabajo para temas específicos y un Equipo Coordinador (EC-SiB) que cumple las funciones de secretaría técnica, acogiendo e implementando las recomendaciones del CD-SiB.

El SiB Colombia promueve la participación activa del gobierno, la academia, el sector productivo y la sociedad civil para lograr la consolidación de información confiable y oportuna que apoye la toma de decisiones a nivel nacional e internacional. Es además, el nodo oficial del país en la infraestructura mundial de información en biodiversidad -GBIF-.

La implementación del SiB Colombia, a partir del 2000, constituyó el primer resultado del nuevo enfoque de gestión de datos e información en el ámbito nacional y se encuentra articulado con el Sistema de Información Ambiental de Colombia (SIAC) como el subsistema de información que soporta el componente de biodiversidad.

# INTRODUCCIÓN

Los datos primarios sobre biodiversidad son la materia prima para la toma de decisiones sobre el manejo de los recursos biológicos. Sin embargo, muchas veces no cuentan con la calidad necesaria para ser utilizados. Para lograr que estos datos sean un insumo confiable de uso en investigación, educación y toma de decisiones, el SiB Colombia ha generado rutinas de calidad de datos implementando herramientas informáticas libres, gratuitas y fáciles de utilizar. Las rutinas están desarrolladas en el lenguaje de programación GREL[1] (General Refine Expression Language) en el entorno del software de código abierto OpenRefine[2], y funcionan como módulos independientes que permiten crear flujos de trabajo específicos para cada conjunto de datos (Registros, Listas, Eventos), integrando otras herramientas de la red de GBIF[3], Canadensys[4] y GeoNames[5], principalmente a través del uso de los servicios web que disponen.

Los procesos de validación y limpieza son claves para maximizar el uso de los datos primarios sobre biodiversidad y son una necesidad para los sistemas de información nacionales sobre biodiversidad como lo es el SiB Colombia y los demás nodos de la red de GBIF. Las herramientas actuales de manejo, validación y limpieza de datos son muy diversas, pero las podemos categorizar en dos tipos: 1) herramientas en línea intuitivas y de fácil uso por cualquier tipo de usuario, o 2) herramientas/programas especializados que requieren unas habilidades o conocimientos avanzados por parte del usuario. El primer tipo está mediado por archivos de textos y requiere múltiples pasos de copiado y pegado que incrementan la probabilidad de error, mientras que el segundo tipo de herramientas aunque tiene un flujo de trabajo estructurado y menor probabilidad de error, requiere una curva de aprendizaje muy alta. En ese sentido, resulta relevante contar con **una herramienta robusta, y fácil de usar, que cubra las necesidades y capacidades de los usuarios interesados en mejorar la calidad de sus datos y que, desde los nodos de GBIF, permita hacer validaciones rápidas sobre las tres dimensiones de los datos sobre biodiversidad: taxonomía, geografía y temporalidad.**

Las rutinas están disponibles en el repositorio de github *'Rutinas de calidad de datos sobre biodiversidad en Open Refine'*, en la Figura 1 se presenta el esquema general de funcionamiento de las rutinas.

[1] https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language
[2] http://openrefine.org/
[3] https://www.gbif.org/
[4] https://www.canadensys.net/
[5] https://www.geonames.org/

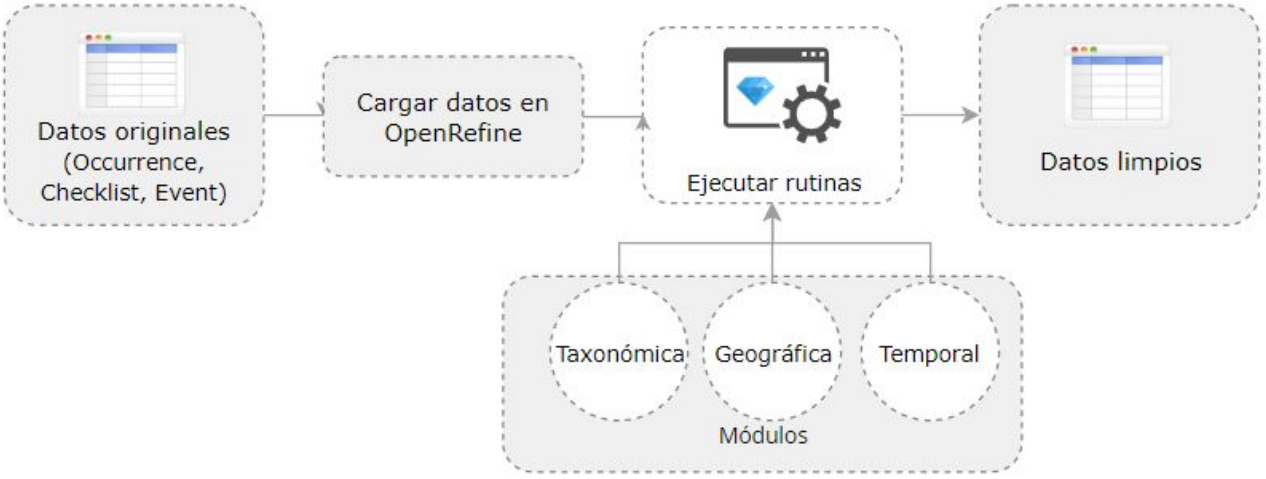



*Figura 1*. Funcionamiento básico de las rutinas en OpenRefine.

Esta guía presenta instrucciones para ejecutar las rutinas y asume un conocimiento básico sobre el manejo de OpenRefine, el cual se puede adquirir rápidamente en documentos como la Guía básica de uso de OpenRefine[6] del SiB Colombia (disponible en ES) o en la wiki de OpenRefine en GitHub[7] (disponible en EN). De igual manera se recomienda tener conocimiento sobre el estándar Darwin Core. Para obtener un contexto general del estándar se recomienda revisar los documentos: '¿Qué es Darwin Core y por qué es importante?'[8] y la '*Guía de referencia rápida*'[9].

# RUTINAS DE VALIDACIÓN

Las rutinas son archivos de texto plano que contienen instrucciones en el lenguaje de programación GREL y son ejecutadas sobre conjuntos de datos cargados en Open Refine.

Para su ejecución se deben tener en cuenta los siguientes requisitos:

- Contar con una instalación de OpenRefine versión 2.6 o superior.
- Los conjuntos de datos deben estar estructurados en el estándar Darwin Core.
- Conexión a internet.
- Para algunas rutinas es necesario cargar archivos adicionales en OpenRefine (disponibles en el repositorio en GitHub).

Para trabajar con conjuntos de datos de más de 1000 registros, se recomienda expandir la memoria de trabajo de OpenRefine[10].

[6] http://repository.humboldt.org.co/handle/20.500.11761/35348
[7] https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions
[8] https://www.gbif.org/es/darwin-core
[9] http://rs.tdwg.org/dwc/terms/
[10] https://github.com/OpenRefine/OpenRefine/wiki/FAQ:-Allocate-More-Memory

Las rutinas están dispuestas en un repositorio abierto en GitHub *'Rutinas de calidad de datos sobre biodiversidad en Open Refine'*[11], con el fin de poder centralizar la documentación de cambios, mejoras y reportes de problemas en su ejecución. Los detalles sobre el funcionamiento de cada rutina están documentados junto con el código de cada rutina para facilitar su uso y minimizar errores de ejecución.

**¿Cómo funcionan las rutinas?**

Las rutinas contrastan la información documentada en el conjunto de datos contra diferentes fuentes de referencia, y se generan elementos/columnas de validación donde se puede identificar la correspondencia entre el archivo original y la fuente de referencia a través de operadores lógicos que generan unos (1) y ceros (0) como indicadores de validación (Figura 2.).

Los **indicadores de validación** se interpretan así:

**(0):** El valor documentado en el conjunto de datos NO coincide con la fuente de referencia, el valor debe ser revisado y ajustado en caso de ser necesario.

**(1)**: El valor documentado en el conjunto de datos coincide con la fuente de referencia, no es necesario tomar acciones adicionales.

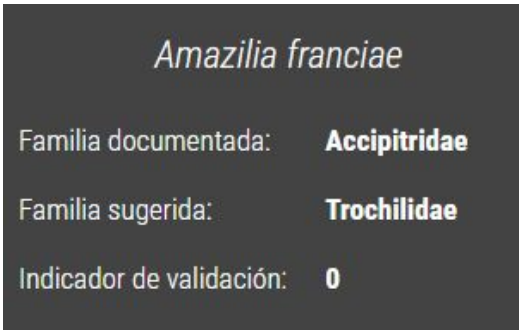



*Figura 2*. Ejemplo del uso de los indicadores de validación para la revisión de los datos.

Las rutinas del SiB Colombia utilizan como fuentes de validación (1) API's (Interfaces de Programación de Aplicaciones) de repositorios globales taxonómicos y geográficos; y (2) archivos de texto plano obtenidos como resultado de herramientas de validación externas y fuentes nacionales oficiales. Se explica a continuación cada fuente:

(1) Las rutinas cuya fuente de referencia es un API, hacen una **consulta al servicio en línea** y extraen la información necesaria para hacer la validación, esta se obtiene en formato JSON y es interpretada por la rutina para hacer la información legible dentro del conjunto de datos. Posteriormente el resultado de la consulta al API es comparado con el valor documentado en el conjunto de datos y se generan nuevas columnas con los indicadores de la validación.

(2) Las rutinas que **usan como fuente archivos de texto plano**, hacen una consulta sobre un archivo cargado previamente en OpenRefine, que posteriormente es comparado con el valor documentado en el conjunto de datos, como resultado de la comparación se generan nuevas columnas con los indicadores de la validación.

[11] https://github.com/SIB-Colombia/data-quality-open-refine

A la fecha se han elaborado seis (6) rutinas de limpieza (Tabla 1), teniendo en cuenta diferentes escenarios al momento de realizar la validación, como la caída temporal de servicios web, o requisitos adicionales según la naturaleza de los datos, como por ejemplo en los datos marinos. Cada rutina cuenta con una descripción en la sección inicial que detalla los requerimientos para su ejecución y su funcionamiento en los idiomas inglés y español.

Tabla 1. Lista de rutinas para la validación de datos primarios sobre biodiversidad

(*) Require acceso a internet para hacer la petición a una API
(~)Requiere que un archivo sea previamente cargado en OpenRefine para la ejecución de la rutina.

| | |
|---|---|
| 1. Validación taxonómica con el API de GBIF(*) | ValTaxonomicAPIGBIF_ValTaxonomicaAPIGBIF.txt |
| 2. Validación taxonómica con Species Match de GBIF(~) | ValTaxonomicSpeciesMatchGBIF_ValTaxonomicaSpeciesMatchGBIF.txt |
| 3. Validación taxonómica con el API de WoRMS (World Register of Marine Species)(*) | ValTaxonomicAPIWoRMS_ValTaxonomicaAPIWoRMS.txt |
| 4. Validación de nombres geográficos(~) | ValNamesGeo_ValNombresGeo.txt |
| 5. Transformación de fechas al estándar ISO con el servicio de conversión de 'Canadensys'(*) | DateTransform_TransformFechas.txt |
| 6. Validación de elevaciones con GeoNames.(*) | ValElevationAPIGeoNames_ValElevacionAPIGeoNames.txt |

## 1. Validación taxonómica con el API de GBIF

Enlace al repositorio:
https://github.com/SIB-Colombia/data-quality-open-refine/blob/master/ValTaxonomicAPIGBIF_ValTaxonomicaAPIGBIF.txt

Obtiene y valida la información taxonómica de un conjunto de datos usando como referencia el árbol taxonómico de GBIF, esto se hace a través de un llamado al API de GBIF[12] basado en los elementos del estándar Darwin Core de nombre científico ('scientificName') y reino ('kingdom') documentados en el conjunto de datos. Como resultado, el llamado retorna la taxonomía superior, nombres aceptados, estatus taxonómico y autoría del nombre científico de acuerdo al árbol taxonómico de GBIF; y la rutina toma los valores del llamado y los compara con los elementos documentados en el archivo base, generando los indicadores de validación.

[12] https://www.gbif.org/developer/species

El llamado al API permite hacer una consulta sobre un número ilimitado de registros, sin embargo, se recomienda ejecutar la rutina haciendo un filtro por nombres científicos únicos, lo cual disminuirá el tiempo de respuesta y agilizará la ejecución de la rutina.

## 2. Validación taxonómica con Species Matching de GBIF

Enlace al repositorio:
https://github.com/SIB-Colombia/data-quality-open-refine/blob/master/ValTaxonomicSpeciesMatchGBIF_ValTaxonomicaSpeciesMatchGBIF.txt

Obtiene y valida la información taxonómica de un conjunto de datos con el árbol taxonómico de GBIF a partir de un archivo de texto plano obtenido de la herramienta en línea de GBIF Species matching[13] y cargado en OpenRefine. La rutina retorna la taxonomía superior, nombres aceptados, estatus taxonómico y autoría del nombre científico de acuerdo al árbol taxonómico de GBIF y los compara con los elementos documentados en el archivo base, generando los indicadores de validación.

Al usar GBIF Species matching como fuente de referencia, el usuario puede realizar una validación previa a OpenRefine directamente en *species-Matching,* la cual es especialmente útil para verificar y resolver sinonimias complejas, como es el caso de los homónimos.

A diferencia del API de GBIF, *species-Match* tiene un límite de consulta de 6.000 registros o nombres científicos. Para evitar exceder el límite de consulta, se recomienda hacer la consulta en *species-Match* por nombres científicos únicos.

## 3. Validación taxonómica con el API de WoRMS (World Register of Marine Species)

Enlace al repositorio:
https://github.com/SIB-Colombia/data-quality-open-refine/blob/master/ValTaxonomicAPIWoRMS_ValTaxonomicaAPIWoRMS.txt

Esta rutina está diseñada especialmente para ser implementada en conjuntos de datos de grupos biológicos marinos, empleando una fuente de referencia específica para estos organismos, así mismo está pensada para que los conjuntos de datos cumplan con los requisitos necesarios para ser integrados en portales de datos de biodiversidad globales: tanto GBIF cómo OBIS (Ocean Biogeographic Information System[14]).

Obtiene y valida la información taxonómica de un conjunto de datos usando como referencia el árbol taxonómico de LifeWatch (LW-TaxBB[15]), esto se hace a través de un llamado al API de WoRMS[16]

[13] https://www.gbif.org/tools/species-lookup
[14] https://obis.org/
[15] http://www.lifewatch.be/en/taxonomic_backbone
[16] http://www.marinespecies.org/rest/

basado en el elemento nombre científico ('scientificName') del estándar Darwin Core documentado en el conjunto de datos. Como resultado, el llamado retorna la taxonomía superior, nombres aceptados, estatus taxonómico, autoría del nombre científico y otros elementos obligatorios para la publicación de datos a través de la plataforma de OBIS, como el identificador del nombre científico de acuerdo a Aphia[17] ('scientificNameID'). La rutina compara los elementos documentados en el archivo base con los retornados por el API, generando indicadores de validación. La rutina permite también obtener información sobre el tipo de hábitat del taxón (Elementos del estándar Darwin Core: *isMarine, isFreshwater; isBrackish, isTerrestial*).

## 4. Validación de nombres geográficos

Enlace al repositorio:
https://github.com/SIB-Colombia/data-quality-open-refine/blob/master/ValNamesGeo_ValNombresGeo.txt

Enlace a archivo división político administrativa oficial de Colombia:
https://github.com/SIB-Colombia/data-quality-open-refine/blob/master/DIVIPOLA_20190417.zip

Desarrollada para estandarizar los contenidos de los elementos de la geografía superior, especialmente *stateProvince, county* y *municipality*, de acuerdo a una fuente de referencia nacional. La rutina contrasta los valores documentados con la información oficial para el país, a partir de un archivo de referencia previamente cargado en OpenRefine, y genera indicadores de validación. Los indicadores permiten identificar dos tipos de errores en la geografía superior; 1) errores de tipeo y gramática y 2) errores de consistencia relacionados con la correspondencia entre entidades geográficas, como municipios (*county*), o centros poblados (*municipality*) que no pertenecen al departamento *(stateProvince*).

El archivo oficial de referencia disponible en el repositorio es generado con la información geográfica para Colombia suministrada por la División Político Administrativa definida por el DANE (Divipola[18]). Vale la pena precisar que esta rutina puede implementarse para otros países, empleando la misma estructura del archivo de la división político administrativa oficial de Colombia, pero con la información geográfica oficial del país de interés.

## 5. Transformación de fechas al estándar ISO con el servicio de conversión de 'Canadensys'

Enlace al repositorio:
https://github.com/SIB-Colombia/data-quality-open-refine/blob/master/DateTransform_TransformFechas.txt

[17] http://www.marinespecies.org/about.php#what_is_aphia
[18] https://geoportal.dane.gov.co/pruebadivipola/

A partir de las fechas documentadas en el conjunto de datos, se realiza una petición al API de Canandensys[19], el cual transforma las fechas en el estándar ISO 8106, retornando también los elementos *year, month,* y *day*. Si algún registro no tiene datos de fecha, la rutina mantiene los elementos *eventDate, year, month* y *day* vacíos. Los formatos de fecha aceptados por el API son:

- Jun 13, 2008
- 15 Jan 2011
- 2009 IV 02
- 2 VII 1986
- 1999/02/24
- 02/17/1921

Hay que tener en cuenta que las fechas con meses en español (enero, junio, etc.), no son convertidas aún por la rutina.

6. **Validación de elevaciones con GeoNames.**

Enlace al repositorio:
https://github.com/SIB-Colombia/data-quality-open-refine/blob/master/ValElevationAPIGeoNames_ValElevacionAPIGeoNames.txt

Realiza un llamado al API de GeoNames[20] (servicio SRTM-1) a partir de los elementos Darwin Core de latitud ('*decimalLatitude*') y longitud ('*decimalLongitude*') en grados decimales y retorna la elevación con una resolución de 30 metros por pixel y la compara con los elementos documentados en el archivo base, generando los indicadores de validación.

# INSTRUCCIONES DE USO

Todas las rutinas se ejecutan de manera similar, los requerimientos específicos se encuentran documentados en cada rutina. En esta sección se presentan instrucciones generales para su ejecución en OpenRefine:

**Paso 1**. Cargue el conjunto de datos que desea validar en OpenRefine, si tiene dudas sobre cómo hacerlo puede consultar la Guía básica de uso de OpenRefine[21]. Para que las rutinas funcionen correctamente el conjunto de datos debe estar estructurado en el estándar Darwin Core.

**Paso 2**. Diríjase al repositorio data-quality-open-refine[22] y seleccione la rutina de validación que desea implementar para su conjunto de datos (Tabla 1).

[19] https://data.canadensys.net/tools/api?lang=en
[20] http://www.geonames.org/export/web-services.html
[21] http://repository.humboldt.org.co/handle/20.500.11761/35348
[22] https://github.com/SIB-Colombia/data-quality-open-refine

**Paso 3**. Para las rutinas que requieren otros archivos cargados previamente en OpenRefine (ver Tabla 1.), cargue los archivos adicionales en OpenRefine; de lo contrario vaya directamente al **paso 4**.

**Paso 4 .** Revise las instrucciones y requerimientos al inicio de cada rutina y verifique que cumple todas las condiciones para la ejecución de la rutina.

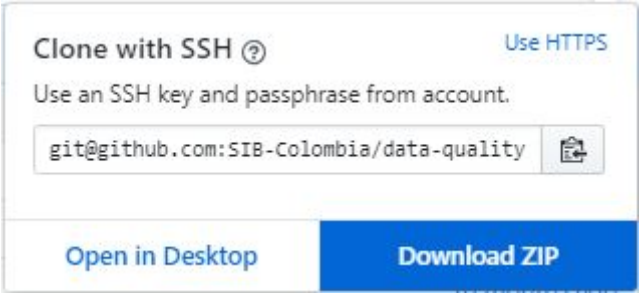
**Paso 5**. Copie el texto de la rutina de validación seleccionada desde el punto señalado. Asegúrese de seleccionar todos los corchetes iniciales ({) o finales (}), esto puede generar errores en la ejecución.

```
Los nuevos datos seran guardados en columnas el inicio del conjunto de datos
Los elementos taxonómicos son reorganizados para facilitar la validación taxonómicas

---Do not copy /No Copiar
--------------------------------------------------------------------------------
Copy from here/ Copiar Desde Aquí
[
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column scientificName using expression grel:value.trim().replace(/\\u00A0/,' ').replace(/\\s+
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
```

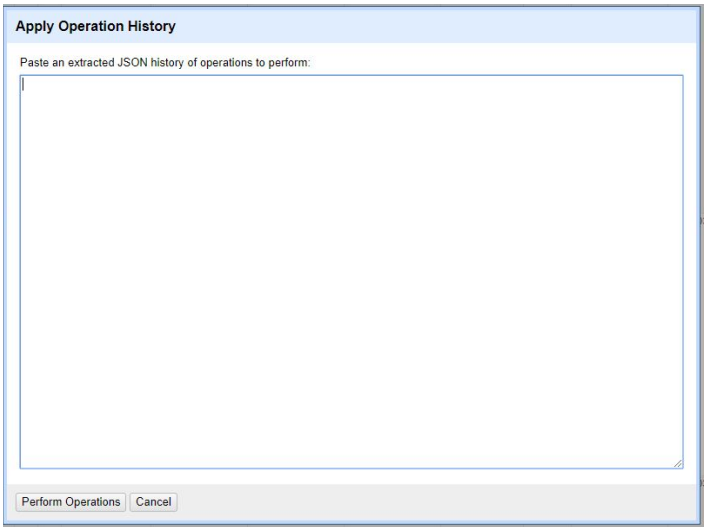
Nota:

- Puede descargar todas las rutinas de validación en su equipo en un archivo comprimido .zip. desde el repositorio.
- Si es usuario de Git (u otro sistema de versionamiento) puede clonar el repositorio en su equipo personal, usando esta guía sencilla[23] de Git. Esto le permitirá tener acceso a las rutinas y actualizar los cambios en su equipo de forma automática.

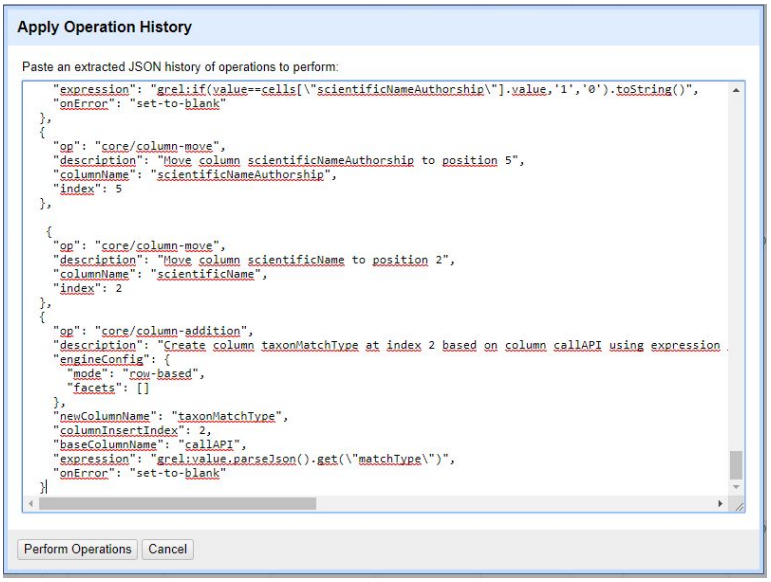



**Paso 6**. Ubíquese en el conjunto de datos a validar en OpenRefine (**paso 1**), diríjase al menú lateral izquierdo, seleccione la opción "**Undo / Redo**" y luego de clic en "**Apply...**". A continuación se abrirá una ventana de texto vacía.

[23] https://rogerdudler.github.io/git-guide/index.es.html



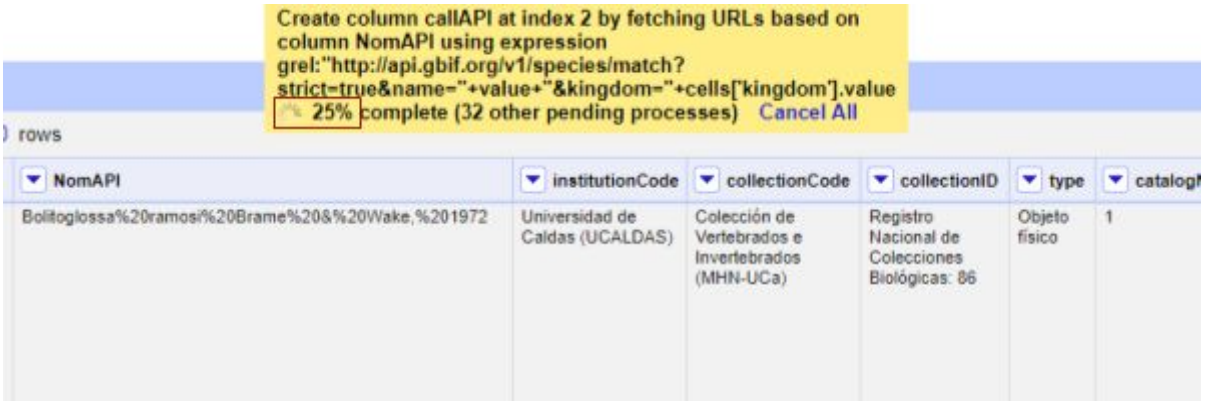

**Paso 7**. Pegue la rutina que había copiado en el **paso 5** en la ventana "**Apply Operation History**" del **paso 6**, y de clic en "**Perform Operations**".




.

**Paso 8**. Espere a que finalice la ejecución de la rutina. Las rutinas que requieren hacer llamados a servicios web, dependen de la conexión a internet, estas consultas toman un tiempo en correr que varía según el número de registros del conjunto de datos, de la velocidad de la conexión y de la memoria RAM del equipo.

El avance del llamado al API se observa en  en la parte superior  de la pantalla.




**Paso 9**. Al terminar la ejecución de la rutina, nuevas columnas aparecerán en el conjunto de datos, estas columnas no pertenecen al estándar Darwin Core y puede identificarlas por su terminación:

- **Suggested**: valores sugeridos resultantes de la validación con las fuentes de referencia, dependiendo de la rutina seleccionada pueden ser sugerencias taxonómicas o geográficas.
- **Validation**: corresponden a los indicadores de validación (unos y ceros) que permiten rastrear diferencias entre el valor original y el valor sugerido, y realizar posteriormente una limpieza de los datos.

**Paso 10**. A partir de las nuevas columnas de validación seleccione los registros donde el valor original y el valor sugerido son diferentes (Identificador de validación = 0) y realice los ajustes que considere necesarios sobre los elementos del estándar Darwin Core. Se recomienda realizar este proceso de limpieza utilizando las funcionalidades de OpenRefine, para ello se sugiere revisar la Guía básica de uso de OpenRefine[24].

**Paso 11**. Una vez terminada la validación y limpieza de sus datos, puede eliminar las columnas resultantes de la validación y dejar solo los elementos Darwin Core.

[24] http://repository.humboldt.org.co/handle/20.500.11761/35348