Recursos publicados a través del SiB Colombia: una contribución a la evaluación de su calidad y a sus procesos de mejora

Estudiante
David Mateo Murillo-Cardona

Directora
Andrea Carolina Castro Moreno

Codirector
Pablo Andrés Guzmán González

Trabajo de Grado
En la modalidad de *Pasantía*

Programa de Biología
Universidad CES
Medellín
Junio 2024

06 de junio de 2024.

Se informa que el estudiante **David Mateo Murillo Cardona** identificado con cédula: No. 1000596882 ha concluido de manera satisfactoria su trabajo de grado titulado "**Recursos publicados a través del SiB Colombia: una contribución a la evaluación de su calidad y a sus procesos de mejora**" en la modalidad de *Pasantía*.

En calidad de directores del trabajo de grado en mención, y luego de haber revisado con detalle y alto rigor científico y académico el presente documento final, se aprueba este Trabajo de Grado como requisito parcial para optar al título de **Biólogo**.

Andrea Carolina Castro Moreno
Cédula: 1032405401
Investigadora – Gerencia de Información Científica
Instituto Humboldt (IAvH)

Pablo Andrés Guzmán González
Cédula: 94331882
Docente – Programa de Biología
Universidad CES

Recursos publicados a través del SiB Colombia: una contribución a la evaluación de su calidad y a sus procesos de mejora

David Mateo Murillo-Cardona

Resumen

Los datos primarios y sus metadatos son la base de muchas actividades alrededor de la biodiversidad. En los últimos 30 años, los avances tecnológicos y los cambios en la política de uso de datos han dado origen al concepto de ciencia abierta, el cual representa poder acceder a los datos sin ningún tipo de barrera, y poder reutilizarlos. Esto es importante para tener respuestas informadas, pero también se deben hacer esfuerzos en garantizar que los datos y sus metadatos sean confiables, íntegros y de calidad. Por esto son relevantes los esfuerzos del Sistema de Información sobre Biodiversidad de Colombia (SiB Colombia), en donde se realizó una pasantía como trabajo de grado, por un acuerdo de cinco meses. Durante este periodo, se apoyaron y optimizaron los procesos de calidad asociados a sus datos publicados. Esto se hizo desde el ajuste de datos y metadatos de recursos asociados a permisos de recolección y colecciones biológicas, y mejorando el script en Python que evalúa la calidad de recursos publicados. Como frutos principales de estas actividades, se eliminó de manera automatizada una barrera técnica que impedía el acceso a los recursos provenientes de permisos de recolección; se mejoró la visibilidad internacional de algunas colecciones biológicas del país a través de una herramienta de alcance global; se mejoró la eficiencia del script que hace diagnósticos de calidad; y se identificaron los principales retos actuales alrededor del uso y publicación de datos de biodiversidad nacionales. Estos resultados contribuyen a la ciencia abierta del país, y a los procesos realizados por el SiB Colombia.

Palabras clave: Datos abiertos, IPT, GRSciColl, Calidad de datos, Flags & Issues.

TABLA DE CONTENIDO

1. Presentación

En el contexto de la informática de la biodiversidad, un **dato** es la unidad mínima de información que evidencia la presencia o ausencia de un organismo en un espacio-tiempo determinado (SiB Colombia, 2020b). A estos datos también se les da el nombre de *datos primarios* (Soberón y Peterson, 2004), y son altamente importantes, pues son el pilar de muchos análisis de biodiversidad, tienen una gran cantidad de usos, y son la base de diversas actividades investigativas (Faith et al., 2013; La Salle et al., 2016). Por su parte, los **metadatos** pueden entenderse como el contexto de los datos (SiB Colombia, 2020b), pues dan información descriptiva sobre quién, por qué y cómo se tomaron los datos, además del cuándo y dónde (SiB Colombia, 2022b).

Los datos primarios se pueden obtener a partir de muchas fuentes: desde los especímenes que están depositados en una colección biológica, hasta las fotos georreferenciadas tomadas con un celular (GBIF Secretariat, 2022). Esto, junto con los avances tecnológicos y los cambios en las políticas de uso de los datos que han ocurrido en los últimos 30 años (como el avance del internet y la digitalización de la información), han generado una revolución en la forma en que se crea, mantiene, distribuye y usa la información de la biodiversidad (Soberón y Peterson, 2004). Uno de los aspectos de esta revolución es la **ciencia abierta**, la cual se puede definir como "el conocimiento accesible y transparente que se comparte y desarrolla a través de redes colaborativas" (Vicente-Saez y Martinez-Fuentes, 2018, p. 1). Cuando este concepto se aplica al contexto de los datos biológicos, implica que se puede acceder a ellos sin barreras técnicas, económicas o legales, y que estos están disponibles bajo un formato y licencia que es amigable para su reutilización (SiB Colombia, 2021b).

Se considera que compartir de forma abierta los datos de biodiversidad, es una parte importante para poder tener respuestas informadas y a tiempo de los impactos que, como seres humanos, estamos produciendo en el mundo (La Salle et al., 2016). Y que tomar tales decisiones informadas es especialmente importante en países megadiversos (Canhos et al., 2015), como es el caso de Colombia. No obstante, debemos tener presente que el reto actual no solo es aumentar la visibilidad de los datos o facilitar el acceso a los mismos, sino también garantizar que se está accediendo a datos íntegros y confiables (Wheeler et al., 2004). Es por esto que se recomienda que en los países existan marcos de gobernanza de información que, además de compartir información biológica de forma abierta, también se preocupen por la calidad e integridad de sus datos y metadatos, pues así se facilita la identificación, descubrimiento y reutilización de estos (Awada et al., 2022).

Es en este contexto donde destaca la importancia del Sistema de Información sobre Biodiversidad de Colombia -SiB Colombia-, pues esta red nacional, coordinada por el Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, busca brindar acceso abierto a datos e información sobre la biodiversidad del país (Gamboa et al., 2019), acompañar a los publicadores para asegurar que se cumplan con estándares de calidad internacionales (SiB Colombia, 2020a), y promover que muchas audiencias utilicen dicha

información; todo para apoyar la gestión de la biodiversidad del país de forma integral, eficiente y oportuna (Villa et al., 2023).

Desde el 2015, el SiB Colombia cuenta con un programa de pasantías, el cual busca desarrollar en profesionales en formación habilidades relacionadas con la publicación, acceso y uso de datos sobre la biodiversidad del país. El programa está dividido en siete enfoques diferentes, los cuales están agrupados en tres líneas: Administración de Contenidos (AC), Gestión y Cooperación (GC) y Productos y Servicios (PS) (SiB Colombia, 2022a). Se tuvo la oportunidad de participar, entre octubre del 2023 y marzo del 2024, en este programa de pasantías bajo un perfil de la Administración de Contenidos, el cual está enfocado en la Informática de la Biodiversidad. La pasantía tuvo un mes de entrenamiento a cargo del Equipo Coordinador del SiB Colombia, y en el resto de meses se realizaron diferentes actividades. En este documento se presenta una reseña del SiB Colombia, así como el contexto, procedimiento y resultados de las actividades realizadas durante la pasantía. Dado que la temática que se trata a continuación tiene algunos términos muy específicos, se puede consultar el **Anexo 9.1**, el cual contiene un glosario que busca ayudar a quienes no estén relacionados con esta temática.

2. Reseña de la institución

El SiB Colombia es la red nacional de datos abiertos sobre biodiversidad, y el nodo oficial del país ante el **GBIF**, o Infraestructura Global de Información en Biodiversidad (SiB Colombia, 2021a). Esta red tiene su origen en el Decreto 1603 de 1994, en el cual se establecen tres de los institutos de investigación que hacen parte del Sistema Nacional Ambiental (Ministerio del Medio Ambiente, 1994), y a partir de su implementación en el 2000, está articulado con el Sistema de Información Ambiental de Colombia (SIAC), como el subsistema de información que soporta el componente de biodiversidad (Escobar et al., 2016). El SiB Colombia está conformado por un Comité Directivo, el cual sienta las bases y traza las directrices de la red; un Comité Técnico, el cual identifica y evalúa estrategias para implementar el plan de acción; y un Equipo Coordinador (**EC-SiB**), el cual implementa las recomendaciones del Comité Directivo y ejecuta el plan de acción (Gómez-Ahumada, 2013).

Desde el 2018, el SiB Colombia ha avanzado alrededor de cinco ejes de trabajo: consolidar la red nacional del SiB Colombia, para que todos cuenten con la capacidad de compartir y usar los datos; proveer datos e información relevante, para cumplir con los requerimientos de procesos de investigación y educación; mejorar la infraestructura informática, para respaldar la integración de la información; llenar vacíos; y mejorar la calidad de los datos e información, para que sean un buen punto de referencia (Escobar et al., 2018).

Gracias a estos esfuerzos, el SiB Colombia registró en enero del 2024 un acumulado de 215 publicadores asociados (principalmente Organizaciones No Gubernamentales, empresas, el sector académico, aunque también autoridades ambientales, centros de investigación,

entre otros); contó con 2850 conjuntos de datos, con más de 28 millones de registros biológicos; y registró unas 1800 personas utilizando sus portales al mes, y sus datos publicados se han usado en más de 41 mil revistas indexadas (Ortiz et al., 2024).

3. Objetivos

3.1 Objetivo general

Apoyar y optimizar los procesos de calidad asociados a la publicación de datos a través del SiB Colombia.

3.2 Objetivos específicos

- Apoyar la aplicación de ajustes en datos y metadatos, en al menos 50 recursos publicados a través del SiB Colombia que presentan problemas de calidad.

- Apoyar la mejora del script en lenguaje Python, que sirve como diagnóstico de la calidad de los datos publicados a través del SiB Colombia, a partir de elementos priorizados en datos y metadatos.

4. Logros alcanzados

- Adquisición de conocimiento teórico-práctico sobre la validación geográfica y taxonómica de datos, su publicación a través del SiB Colombia, y la gestión de bases de datos con registros biológicos utilizando el software OpenRefine y el lenguaje de programación Python.

- Creación de un script en Python para asignar, de forma automatizada, los creadores de recurso en más de 4000 metadatos presentes en el portal IPT_Permisos.

- Diagnóstico del estado de los recursos asociados a las colecciones biológicas del país, con respecto a su indexación en el portal de Registro Global de Colecciones Científicas (GRSciColl por sus siglas en inglés).

- Revisión y mejoramiento de un script en Python que genera diagnósticos de la calidad de los datos publicados a través del SiB Colombia, según falencias encontradas en el registro de elementos del estándar Darwin Core (**DwC**).

- En todos los scripts desarrollados, se implementó el pensamiento programativo que apunta a la reproducibilidad y transparencia del código.

5. Dificultades

Dado que el programa de pasantías es un proceso secuencial, varias de las actividades realizadas tuvieron como base los procedimientos hechos por pasantes anteriores. En algunos casos, se encontró que no había una documentación completa sobre los procedimientos empleados, los resultados obtenidos, o el contexto de la actividad; por lo que se dificultó entender algunas situaciones.

Uno de los principales retos con el script de Python que genera diagnósticos de calidad, es que se necesita manipular dos archivos muy pesados (18 GB y 24 GB). Si bien hay medios por los cuales se pueden compartir sin problemas las versiones comprimidas de estos archivos, los editores de texto comunes no están optimizados para visualizar o gestionar tanta información. Por estas razones, se tuvieron que emplear dos semanas buscando alternativas, tanto de equipo como de software, para poder ejecutar el script.

6. Resultados

6.1 Capacitación en informática de la biodiversidad

Como parte del mes de capacitación, se consultaron diversas fuentes señaladas por el EC-SiB que abordan qué es el SiB Colombia, sus funciones y las diferentes rutas de publicación de los tipos de datos que se gestionan (registros biológicos, eventos de muestreo y listas de especies). También se participó de los laboratorios virtuales[1], los cuales buscan dar fundamentos para la publicación y gestión de datos sobre biodiversidad. En cada caso, se profundizaron los conceptos a través de discusiones dentro del equipo de pasantes.

Como resultado, se obtuvo conocimiento teórico-práctico de las diferentes validaciones que son realizadas al momento de la publicación de un conjunto de datos, como son la validación taxonómica, utilizando algunas herramientas de páginas como GBIF, Servicio de Resolución de Nombres Taxonómicos (TNRS), o Registro Mundial de Especies Marinas (WoRMS); y la validación geográfica, convirtiendo coordenadas y aplicando las rutas de validación del SiB Colombia. También se desarrollaron ejercicios utilizando los programas OpenRefine y QGIS, y se realizaron ejercicios de publicación a través de la herramienta IPT. Finalmente, también se apoyó la identificación de posibles errores en estas guías, los cuales fueron comentados al EC-SiB.

[1] Enlace de los laboratorios virtuales https://biodiversidad.co/formacion/laboratorios

6.2 Apoyo a la migración de recursos al IPT_Permisos

Desde el 2013, por normatividad colombiana, todos los titulares de un Permiso de recolección para elaborar Estudios Ambientales, o un Permiso Marco de Recolección, o de un Permiso Individual de Recolección, deben reportar al SiB Colombia la información asociada a los especímenes recolectados, para así obtener un certificado que debe ser enviado a la autoridad competente (Ministerio de Ambiente y Desarrollo Sostenible, 2013b, art. 6, 2013a, arts. 9, 14, respectivamente). Para dar cumplimiento a esta normatividad, el SiB Colombia habilitó desde julio del 2014 el portal *Certificado de Reporte*, en el cual, a través de unas credenciales, cada titular creaba un recurso para indicar los especímenes recolectados bajo el permiso, y así obtenía el certificado (Amariles et al., 2015).

Este portal presentaba problemas de interoperabilidad y experiencia de usuario, por lo que desde el 2022 se desarrollaron un grupo de mejoras que implicaron la migración de este proceso hacia el IPT_Permisos[2] y a un nuevo portal para la generación del certificado[3] (Villa et al., 2023). El **IPT** (o *Integrated Publishing Toolkit*) es un software de código abierto por medio del cual entidades publicadoras en el SiB Colombia, y los sistemas globales como el Sistema de Información sobre Biodiversidad Oceánica (OBIS), la Red Global de Biodiversidad Genómica (GGBN) y GBIF, crean y comparten conjuntos de datos de biodiversidad; adaptando diferentes instancias de publicación dependiendo del origen de los datos (Marentes et al., 2023).

La plataforma del IPT_Permisos, al ser homóloga a los otros IPT utilizados por el SiB Colombia, soluciona los problemas que se estaban presentando (Villa et al., 2023). Como parte de este proceso, uno de los pasantes diseñó y utilizó una automatización con *UiPath*[4] para migrar al IPT_Permisos aquellos conjuntos de datos que estaban en el antiguo portal, que eran más de 4000 (Díaz et al., 2023). Sin embargo, durante la migración no se asignó en el metadato quién era el titular del recurso, por lo que los titulares no podían encontrar ni gestionar sus recursos en el nuevo IPT. En esta pasantía, se aplicaron los ajustes necesarios en el metadato de 4391 recursos que ya habían sido migrados, dando solución a este problema. A continuación se detallan los resultados de la actividad.

6.2.1 Base de datos Inventario Migraciones

Antes de realizar las asignaciones, primero se necesitaba saber qué recursos fueron migrados al IPT_Permisos y cuál era su titular respectivo. Para esto, primero se depuró y organizó el archivo creado por pasantes anteriores, en el que se tenía registro de los recursos presentes en el portal *Certificado de Reporte*, y su estado de migración al IPT_Permisos. Durante la depuración, se excluyeron recursos que no fueron migrados, se llenó la información que estaba vacía, se ajustaron algunas fechas al formato ISO 8601, se

[2] Enlace del IPT_Permisos https://ipt.biodiversidad.co/permisos/
[3] Enlace del nuevo portal del certificado https://biodiversidad.co/certificados/publicacion-permisos/login/
[4] Enlace UiPath https://www.uipath.com/

corrigieron los enlaces mal asignados, y se señalaron las entradas y páginas duplicadas (Hoja *Migraciones CR-SiB*, **Anexo 9.2**). A cada una de estas situaciones se le asignó un código de color, los cuales están explicados en el archivo depurado (Hoja *metadata*, **Anexo 9.2**).

Como resultado de la depuración, se obtuvieron 4416 recursos, a los cuales se les creó una nueva columna para indicar el correo de su titular respectivo. Para 1901 recursos fue posible relacionar el correo de forma automática, con ayuda del EC-SiB, a partir de la coincidencia exacta con un listado actualizado de los publicadores en el IPT_Permisos. Para los otros 2515 recursos, se organizó una tabla en la cual se indicaron, para cada titular, los cambios efectuados para que coincidiera completamente con alguno de la lista de los publicadores (Hoja *asignación*, **Anexo 9.3**). No fue posible asignarle el correo a 15 de estos recursos, puesto que después de revisarlos puntualmente, no hubo certeza de cuál era el titular respectivo.

6.2.2 Script en Python para automatizar la asignación

Una vez se le asoció un correo a la mayor cantidad de recursos posibles (4401), se buscaron diversas alternativas para asignar de forma automatizada este dato en los metadatos del IPT_Permisos. Se encontró que la librería Selenium de Python permite interactuar con las páginas web, por lo que se buscó documentación en internet y se hizo un diseño tentativo de un script que, a partir de un archivo csv que tuviera los enlaces de los recursos y los correos de sus titulares, ingresara al portal y asignara los correos mediante un ciclo (*loop*).

Para poner a prueba este diseño tentativo del script, se crearon archivos csv con 20 recursos del inventario (se muestra un ejemplo en la hoja *trial_asignación*, **Anexo 9.2**). Con esto, se identificaron dos errores que pueden aparecer durante las asignaciones automatizadas, y se ajustó el script de forma que éste no se detuviera si ocurrían, sino que los imprimiera al final para que se pudieran revisar manualmente. Si bien estos ajustes permiten utilizar el script con archivos csv de cualquier cantidad de recursos, la asignación se hizo en varios momentos para disminuir la cantidad de revisiones puntuales al final de cada ejecución.

Los resultados de la asignación automática se muestran en la hoja *base_asignación* del **Anexo 9.2**. La versión final del script, junto con un video que ejemplifica la ejecución del mismo, y una explicación a detalle de su funcionamiento, se encuentran disponibles en el **Anexo 9.4**. Las asignaciones se realizaron en 10 momentos diferentes, con archivos que variaron entre 24 y 861 recursos. Para cada uno de ellos, se registró el tiempo total empleado en recorrer todo el archivo, y el promedio de la asignación de un solo recurso (**Tabla 1**). Se encontró que el tiempo total de asignación (tiempo_ciclo) fue completamente dependiente a la cantidad de recursos en el archivo (correlación de 0.99). Por otro lado, el tiempo que tardó en asignarse el autor de un solo recurso (tiempo_fila), fue independiente al tamaño del archivo (correlación de 0.14), con una duración promedio de 8.78 (± 0.25) segundos, y un coeficiente de variación del 2.85%. Se presume que esta leve variación podría atribuirse a fluctuaciones en la conexión a internet en los distintos momentos de asignación. Aunque manualmente es posible asignar un correo en menos tiempo, se

considera que esta velocidad es óptima, dado que el proceso manual aumentaría el riesgo de errores. Además, con computadores superiores a 8 GB de RAM se podrían disminuir los tiempos de espera que tiene el script y por tanto se obtendría un mejor desempeño.

Tabla 1. Cantidad de recursos (N) y tiempo en minutos que tomó la asignación (tiempo_ciclo) de los archivos csv utilizados con el script (Prueba), además del tiempo en segundos que en promedio tomó la asignación de un solo recurso (Promedio tiempo_fila). También se muestra por columna el promedio, la desviación estándar (SD) y el Coeficiente de Variación en porcentaje (CV (%)).

Prueba	N	Tiempo_ciclo (min)	Promedio tiempo_fila (s)
1	317	44.61	8.55
2	216	33.61	9.34
3	267	38.11	8.56
4	284	41.39	8.74
5	360	52.81	8.80
6	632	90.72	8.61
7	604	90.92	9.03
8	861	126.54	8.82
9	619	91.18	8.84
10	24	3.39	8.55
Promedio	418.40	61.33	8.78
SD	251.09	36.96	0.25
CV (%)	60.01	60.27	2.85

Después de realizar las asignaciones, se generó una muestra aleatoria de 100 recursos, a los cuales se les verificó que el nombre de quien reportó el recurso (especificado en un párrafo de los metadatos básicos) coincidiera con el correo que fue asignado de forma automatizada (Hoja *evParrafos*, **Anexo 9.2**). Por otro lado, también se hizo una comparación de la lista de publicadores con respecto a la lista que aparece en IPT_Permisos (Hoja *listado*, **Anexo 9.2**), y se notificaron los casos particulares al EC-SiB.

6.3 Diagnóstico de la indexación en GRSciColl

En el 2022, el SiB Colombia comenzó la implementación del Registro Global de Colecciones Científicas, o GRSciColl, por sus siglas en inglés (Villa et al., 2023). Este es un portal de GBIF que contiene información de las colecciones científicas, incluyendo su ubicación, contactos, identificadores y estadísticas de los especímenes contenidos; y que se apoya en los nodos nacionales, en este caso el SiB Colombia, para importar la información de las colecciones que están registradas en el país (GBIF, 2023).

Aunque hay varias colecciones biológicas que han publicado datos a través del SiB Colombia, la información de algunas de ellas no se ve asociada en GRSciColl. Se considera que esto ocurre porque, los datos de los recursos, deben cumplir con las siguientes condiciones particulares para que puedan ser indexados en dicho portal: 1. El elemento *basisOfRecord* debe estar documentado como “PreservedSpecimen” (término que solamente puede ser utilizado en los recursos que son publicados por colecciones

biológicas). 2. Que de los elementos *institutionCode*, *collectionID*, *institutionID* y *collectionCode* haya por lo menos dos que coincidan con la información reportada en GRSciColl.

Con respecto a los recursos publicados a través del SiB Colombia, el *institutionID* y el *collectionCode* no poseen información relacionada con GRSciColl (la primera suele ser el NIT de la entidad, y la segunda suele ser un código interno del SiB). Por tanto, la indexación de un recurso sólo sería posible si hay relación entre *institutionCode* y *collectionID*. Para lograr esto, es necesario que el *institutionCode* coincida con el código que aparece en GRSciColl, y que el *collectionID* tenga el enlace de la colección biológica en el portal de GRSciColl. Por ejemplo, el Herbario Federico Medem tendría el siguiente enlace: https://scientific-collections.gbif.org/collection/2ba1e155-bd1e-4a13-9ca4-069351897604.

Durante esta pasantía se organizó el archivo con las colecciones biológicas publicadoras hasta diciembre del 2022, y se hizo el primer diagnóstico de cuáles de estas aparecen en GRSciColl, así como de cuántos de sus recursos estaban siendo mapeados. Adicionalmente, se aplicó el ajuste en el *collectionID* de 106 conjuntos de datos, logrando la indexación de 36 de ellos; lo cual es una contribución al proceso para mejorar el acceso y visibilidad mundial de las colecciones biológicas del país. A continuación se detallan los resultados de la actividad.

6.3.1 Actualización del inventario

Se utilizó la extensión de Web Scraper[5] para consolidar en una tabla el nombre y el enlace de GRSciColl de las 144 colecciones y 70 instituciones que aparecen para Colombia hasta la fecha (2023-11-17) (Hojas *enlacesCol* y *enlacesIns*, **Anexo 9.5**). Posteriormente, se revisaron los 267 recursos presentes en el inventario de Colecciones, según el corte del 31 de diciembre del 2022 (Hoja C*olecciones20211231*, **Anexo 9.5**). A partir de esto, en algunos recursos se actualizó el número de registros biológicos documentados hasta la fecha (2023-11-24), y en otros se actualizó el nombre del recurso o su enlace. Luego se compararon los nombres de las colecciones que aparecen en el inventario, con los nombres de las colecciones que aparecen en GRSciColl para Colombia. A raíz de esta comparación se identificaron diversas situaciones que debían ser resueltas por el EC-SiB, las cuales fueron notificadas con detalle. Aunque muchas de estas situaciones fueron resueltas, para 21 recursos no se pudo asociar una colección de GRSciColl, ya que requiere ajustes por parte de otras entidades (Hoja *RNCvsGR*, **Anexo 9.5**). Una vez hecho eso, se pueden revisar otra vez estos 21 recursos y determinar a qué colección de GRSciColl pertenece.

[5] Enlace de Web Scraper https://www.webscarper.io/

6.3.2 Análisis de todos los portales de GRSciColl

Luego de haber asociado la mayor cantidad de recursos con una colección en GRSciColl, se quiso determinar, para cada recurso, cuál es la cantidad de registros que aparecen indexados en el portal respectivo. Para esto, se utilizó el filtro *Dataset* (el cual está presente en la sección de Specimens de cada colección de GRSciColl), pues permite saber cuáles son los conjuntos de datos que están siendo indexados en una colección de GRSciColl, junto con la cantidad de registros que están alimentando la tabla.

Con base en esto, se creó un inventario-diagnóstico que muestra, para cada colección biológica, cuáles son los recursos que están siendo indexados en su página GRSciColl y cuáles no; ambos con su cantidad de registros biológicos (Hoja *mapeo RRBB*, **Anexo 9.5**). Si un recurso estaba alojado en el IPT_SiB, pero no estaba siendo indexado en GRSciColl, entonces se modificó el *collectionID* para que tuviera el enlace de la página de GRSciColl respectiva, y días después se observó si hubo cambios en la indexación. Como resultado, se identificaron 363 recursos, los cuales fueron clasificados en 11 categorías (**Tabla 2**). Estas fueron explicadas con más detalle al EC-SiB.

Tabla 2. Categorías resultantes del diagnóstico de la indexación de recursos en GRSciColl. Se especifica el nombre de la categoría (Estado), una descripción de la misma, y la cantidad de recursos asociados (N).

Estado	**Descripción**	N
OK	Recurso de una colección biológica cuyos registros están completamente indexados en GRSciColl sin necesidad de ajustes.	112
Agregar | Ajustado	Recurso de una colección biológica, alojado en el IPT_SiB, y que se le ajustó *collectionID*, pero sigue sin estar indexado en GRSciColl.	68
Faltante	Recurso indexado en GRSciColl, que se considera de una colección biológica, pero que no está copiado en el inventario.	51
Quitar	Recurso que está indexado en GRSciColl, pero se considera que su origen no es de una colección biológica, sino de un proyecto, por lo que debe ser quitado.	47
Agregado	Recurso de una colección biológica, alojado en el IPT_SiB, cuyos registros se indexaron completamente en GRSciColl después de ser ajustado.	36
Por revisar	Recurso al que no se le asignó en el inventario un enlace de GRSciColl, por una razón explicada.	20
Agregar | Ajustar | IPT-CD	Recurso de una colección biológica, alojado en algún IPT de un miembro del Comité Directivo del SiB, y que por tanto no se le pudo ajustar el *collectionID*.	14
?	Recurso al que se le asignó en el inventario un enlace tentativo de GRSciColl, pero se duda que sea el adecuado, y por tanto no fue ajustado.	8
Menos	Recurso de una colección biológica, el cual no necesitó ajuste para su indexación en GRSciColl, pero que allí se muestran menos registros de los que realmente tiene el recurso.	4

? | Ajustado	Se diferencia de los "?" porque al recurso se le aplicó el ajuste en *collectionID*, sin embargo no quedó indexado a GRSciColl.	2
Agregar | previamente ajustado	Recurso de una colección biológica, alojado en el IPT_SiB, al cual alguien del EC-SiB le ajustó el *collectionID*, pero sigue sin indexarse.	1
TOTAL		363

Si se consideran los recursos que ya estaban indexados, junto con los que fueron indexados gracias al ajuste, y los que no habían sido considerados (OK + Agregado + Faltante), se tiene que el 54.8% de los recursos ya no requieren ajustes (**Figura 1**). En el marco de esta pasantía, no fue posible realizar los ajustes en el porcentaje restante, pero se realizó una propuesta de cinco pasos para ir abordando las categorías que todavía no están indexadas. Dicha propuesta fue entregada al EC-SiB, e incluye los casos particulares que fueron encontrados durante el diagnóstico, así como de un mapeo de *institutionCode* y *collectionCode* que se hizo en 118 recursos.

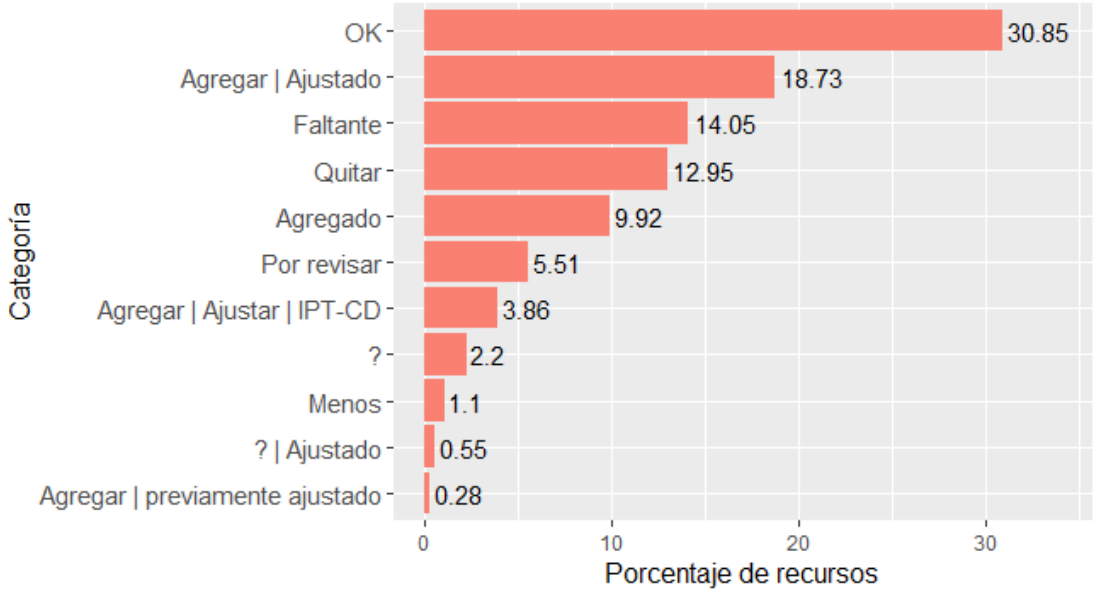

Figura 1. Distribución de los 363 recursos encontrados, de acuerdo con su estado de indexación en la página de GRSciColl de la colección respectiva.

6.4 Modificación del script que genera diagnósticos de calidad

Al momento de publicarse datos en el SiB Colombia, hay dos retos grandes para la calidad de los datos:

1. Actualmente hay más de 200 publicadores asociados (Ortiz et al., 2024), los cuales provienen de diferentes sectores. Esto demuestra que el origen de los datos es muy heterogéneo, y se ha dicho que ante esta situación se deben utilizar

varios procedimientos para detectar y corregir problemas de calidad (Soberón y Peterson, 2004).

2. El constante proceso de revisión y actualización del estándar del DwC (ver fechas de actualización en Wieczorek, 2023), implica que los conjuntos que fueron publicados bajo una versión del estándar, pueden tener problemas de calidad con respecto al estándar actual.

Por estas dos razones, el EC-SiB busca realizar limpiezas retrospectivas para asegurar la calidad de los datos que han sido publicados previamente por el SiB Colombia (Díaz et al., 2022). Esto se viene realizando desde el 2018, y uno de los criterios que se utiliza durante estas limpiezas son elementos del DwC que se han definido como prioritarios (Plata et al., 2022); pero también se pueden utilizar los **Issues & Flags** de GBIF, los cuales son "advertencias" que se agregan a cada registro durante el proceso de indexación de los datos, y que informan sobre problemas de calidad comunes (Buitrago, 2020).

Dado que los métodos para evaluar la calidad de datos son generalmente dispendiosos (Díaz et al., 2019), desde el 2021 se viene desarrollando un script en Python para automatizar la identificación de inconsistencias en conjuntos de datos ya publicados, y hacer más eficiente el proceso (Salinas, 2022). Durante una pasantía anterior, se diseñó un esquema para darle "pesos" (que de aquí en adelante se le denominará **magnitudes**) a algunos Issues & Flags (algunos generados por GBIF, otros creados por el SiB Colombia. Ver **Anexo 9.6.1**), y se desarrolló un script en Python el cual utiliza estas magnitudes para "calificar" los conjuntos de datos publicados, para priorizar cuáles deben ser ajustados primero (**Anexo 9.6.2**).

A partir de ese trabajo surgieron algunas recomendaciones, tales como: 1. Buscar reducir el sesgo relacionado con tener una gran cantidad de registros biológicos. Esto se debe a que, de forma general, los conjuntos más grandes son los que quedaron con "mayores calificaciones", y se consideró que la solución estaría en reevaluar los valores de los "pesos" (magnitudes) (listados en el **Anexo 9.6.3**). 2. Optimizar los procesos para que el análisis se pueda desarrollar en computadores de bajas capacidades, donde se planteó como una opción crear chunks para generar procesos cíclicos que después se unen en un único reporte. 3. Crear más Issues & Flags de otros elementos del DwC (Cano, 2023).

En la presente pasantía se revisó el script, y se propusieron cambios que hacen más homogénea la sintaxis o aclaran algunos de los procedimientos realizados; se sugirieron otros cambios que mejoran o amplían el funcionamiento; se evaluaron diferentes métodos que pueden servir al momento de ejecutar el script en equipos de menos capacidades; y se implementó una nueva forma de utilizar las magnitudes de los errores para calificar los recursos, la cual elimina el sesgo asociado al tamaño del conjunto de datos (sin tener que modificar el esquema de magnitudes creado en la pasantía anterior). Todos los archivos asociados al nuevo script están en el **Anexo 9.7**, y en cada momento que se sugiere un cambio de sintaxis, se tiene documentación de pruebas que lo sustentan (**Anexo 9.7.1**). A continuación se detallan los puntos anteriormente mencionados.

6.4.1 Propuestas para mejorar la claridad

Con respecto a la sintaxis, en todo el script se homogeneizó la forma en que se escriben los comandos *pandas.merge()* y *numpy.where()*; se simplificaron líneas de código cuando fue posible; se omitieron argumentos si no afectaba la funcionalidad o claridad, y se cambiaron los nombres de algunas variables por nombres más descriptivos. Con respecto a aclaraciones, se buscó manejar el mismo estilo de comentarios para explicar el procedimiento del script, incluyendo el funcionamiento de algunos argumentos utilizados.

También, para mejorar la reproducibilidad, se reestructuró la forma en que se gestionan las rutas de los archivos. Anteriormente, se debían manipular 14 líneas de código (a lo largo del script) cada vez que se quería utilizar un computador diferente, o si se modificaban los nombres de los archivos. Con el cambio, solo se modifican cinco líneas, las cuales están al principio del script.

6.4.2 Sugerencias para mejorar el proceso

Algunos de los elementos del DwC deben ser documentados con un vocabulario controlado[6]. Es por esto que varias Issues & Flags (de aquí en adelante **flags**) buscan identificar si se está utilizando dicho vocabulario en los elementos respectivos. En ese sentido, se encontró que el código permitía algunas flexibilidades (por ejemplo, si el vocabulario controlado es "HumanObservation" en el script se aceptaba "theHumanObservation"). Estas líneas se ajustaron para que solamente acepten los valores exactos de dicho vocabulario. Además, la flag relacionada con el elemento *type* se ajustó para que sólo admitiera el vocabulario controlado en inglés, y así cumplir con el lineamiento del estándar Darwin Core.

Se ajustó el código de *flagGEO_textoenCoordenadas* y *flagCollectionID* para hacer más preciso su funcionamiento (el cual está explicado en el **Anexo 9.7.1**), y se propuso la creación de la *flagEnlaceCollectionID*, en la cual, si un registro pertenece a una colección, y no está el enlace de GRSciColl, entonces se identifica como error. Para asignar la magnitud de esta flag, se siguió el mismo procedimiento establecido en la pasantía anterior, y quedó con una magnitud de 11 (**Tabla 3**). Los criterios usados para llegar a dicha magnitud están explicados en el **Anexo 9.7.5**, y dicha flag ya fue agregada al archivo utilizado en el script (**Anexo 9.7.6**).

[6] Listado de los elementos del Darwin Core y los casos en que se utiliza un vocabulario controlado https://biodiversidad.co/elementos-darwin-core

Tabla 3. Valores asignados a la flagEnlaceCollectionID, cuya suma es el valor de su magnitud. Los valores se asignaron de acuerdo con los criterios establecidos en la pasantía anterior. También se especifica el razonamiento bajo el cual se llegó a dicho valor.

Criterio	Valor	Razón
Dificultad de edición a partir de información primaria	4	Porque requiere ir a buscar el enlace y realizar algunas validaciones con el Registro Nacional de Colecciones.
Importancia para publicación	3	Porque es obligatorio solo para los recursos de colecciones, ya que si no se documenta no se indexan los registros en GRSciColl.
Importancia de la validación para asegurar la calidad del registro	3	Porque más que ser un vocabulario controlado, es un estándar sugerido.
Importancia de acuerdo a la frecuencia de inconsistencia en la documentación del elemento	1	Porque a partir del 2022 ya se volvió estándar la implementación del enlace. Por lo que aplicaría para los recursos antiguos, que se sabe son aproximadamente la mitad.

6.4.3 Evaluación de métodos para equipos de bajos recursos

Primero, se debe considerar que dos archivos que se importan en el script son muy pesados (18 GB y 24 GB), y estos se usan para unificar algunas de sus columnas en una sola tabla. Aunque en el script se importan solo unas cuantas columnas (26 de 181, y 16 de 69, respectivamente), cuando se intentó correr el script en un computador de 8 GB de RAM y un procesador de 2 GHz, éste se congeló.

Desde el SiB Colombia se facilitó el acceso, mediante AnyDesk (versión 7.0.15), a un computador con 16 GB de RAM y procesador 3.40 GHz. Gracias a esto, se pudo superar la barrera inicial relacionada con la importación de estos archivos. En ese equipo se utilizaron dos interfaces de usuario para Python: RStudio (versión 1.4.1717) con el paquete reticulate (versión 1.34.0), y Spyder (versión 5.4.3). Con ambas se midió el tiempo que le toma importar el mismo archivo, con tres procedimientos distintos, y se encontró que la interfaz de Spyder siempre obtuvo mejores resultados (por ejemplo, 187.12s en Spyder, comparado con 240.98s con reticulate. **Anexo 9.7.2**). Sin embargo, para darle más soporte a estos resultados, sería necesario ejecutar los comandos varias veces (por ejemplo 10), y tomar el valor promedio con su desviación estándar. Se recomienda hacer esto en futuros casos.

Con Spyder también se probaron diferentes metodologías para importar los archivos más pesados (incluyendo la lectura por chunks, ya que fue sugerido en la pasantía anterior). Con cada una se midió el tiempo antes y después de la ejecución del comando, se registró el resultado, y por medio de un comando se limpió todo el espacio posible de la RAM antes de continuar con el próximo comando. En esta comparativa, guardar e importar el archivo en formato feather con compresión zstd obtuvo los mejores resultados (**Anexo 9.7.3**), con la ventaja adicional de que dichos archivos son menos pesados; y en segundo lugar quedó la metodología de utilizar *dtypes*. Sin embargo, debe considerarse que cuando se utiliza el

formato feather, se realiza internamente una compresión que puede generar problemas de incompatibilidad con algunas columnas. En ese caso, se recomienda usar dtypes.

Como propuesta final para los equipos de bajos recursos, se compartimentalizó el script. Es decir, ya no se tienen que importar los archivos más grandes cada vez que se corre el script, sino que el archivo se dividió en dos: uno principal, en el cual se importa una tabla ya depurada (en formato feather); y el secundario, en el cual se hace la gestión de los archivos pesados, y tiene como resultado el archivo utilizado por el script principal (**Anexo 9.7.4**). De esta forma, si se quiere cambiar un criterio de la evaluación (por ejemplo, crear otra flag), no se tienen que ejecutar los comandos iniciales que requieren tanto procesamiento, sino que se parte desde un punto que es más ágil. Además, en el script principal se incluyó un método opcional para hacer el análisis del script con un subconjunto aleatorio de recursos, lo que permite mayor fluidez durante el análisis y un menor requerimiento de procesamiento.

6.4.4 Implementación nuevo método para asignar magnitudes

En Cano (2023) se diseñó un método para darle un "peso" (magnitud) a 44 flags seleccionadas (es decir, no todas las flags existentes reciben una magnitud). Estas magnitudes oscilan entre 7 y 16, donde los valores más altos, representan una mayor importancia de la flag. Estos valores tienen una frecuencia mínima de uno y máximo de nueve (**Figura 2**).

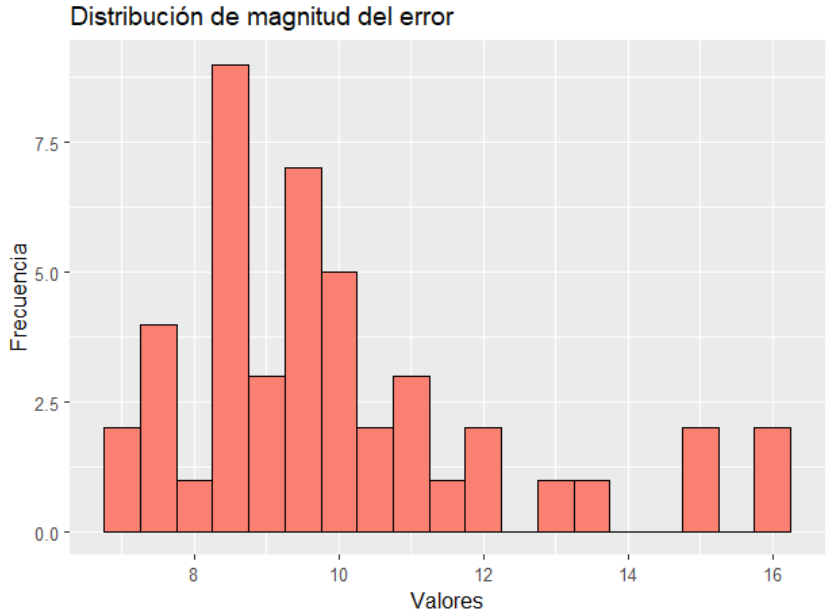

Figura 2. Histograma de los valores (magnitudes) de las 44 flags priorizadas en la pasantía anterior.

En el script desarrollado en la pasantía anterior, estas magnitudes son aplicadas a los recursos de la siguiente forma:

1. Asumamos que un recurso tiene 100 registros.
2. Un solo registro de ese recurso tiene 4 errores (flags 1, 2, 6 y 8), cada error con su magnitud **f** (supongamos, 13, 8, 16 y 9.5). El valor total para ese registro sería **E1**. Por otro lado, otro registro tiene 2 errores, correspondientes a la *flag1* y *flag3* (supongamos que esta flag tiene una magnitud de 9). El valor de ese registro es **E2**.

$$E_1 = f_1 + f_2 + f_6 + f_8$$
$$E_1 = 13 + 8 + 16 + 9.5$$

$$E_2 = f_1 + f_3$$
$$E_2 = 13 + 9$$

3. Pero dentro del recurso de 100 registros, hay en total 25 registros que presentan *mínimo* una flag (recordar que de esos 25 registros, las E pueden ser diferentes). El valor total para ese recurso sería R1.

$$R_1 = E_1 + E_2 + \cdots + E_{25}$$

4. Ese valor R es el que se compara, y determina qué tan importante es hacer los ajustes para uno u otro recurso.

En otras palabras, para cada recurso (R), se suman todos los errores de todos sus registros. Bajo este tipo de procedimiento, las calificaciones más altas están asociadas a los recursos más grandes (más cantidad de registros, lo que representa mayores cantidades de E), y por tanto esos son los que quedan posicionados como los primeros para ser revisados.

Para solucionar este problema asociado al tamaño del recurso, se planteó hacer un promedio ponderado de las magnitudes, es decir:

1. Si se tiene un recurso, con un total de 100 registros, y de ellos 20 registros tienen la *flag1*, 30 tienen la *flag2*, 10 tienen la *flag6* y 50 tienen la *flag8*, entonces las proporciones de esas flags son 0.2, 0.3, 0.1 y 0.5 respectivamente. A estas proporciones se les conoce como **pesos**.

2. Dado que las magnitudes **f** de esas flag son 13, 8, 16 y 9.5, respectivamente, entonces tendríamos el promedio ponderado W para ese recurso:

$$W = \frac{(0.2 * 13) + (0.3 * 8) + (0.1 * 16) + (0.5 * 9.5)}{(0.2 + 0.3 + 0.1 + 0.5)}$$

   Es decir, se está sumando cada (peso * magnitud), y se divide por la suma de los pesos.

De esta forma, se elimina la importancia asociada al tamaño de los recursos (puesto que ya queda en términos de proporción y no de cantidad de registros) y tiene la ventaja de que se tiene la certeza de su significado: si el resultado es 10.31 (como en este caso), esto representa que la mayoría de los errores que tiene el recurso, tienen una magnitud alrededor de 10.

Se hizo una comparación, con 315 recursos aleatorios, de la estrategia empleada en el script anterior con respecto a esta estrategia (**Figura 3**), y es evidente que en el método anterior los resultados tienden entre 0 y el infinito, mientras que ahora los resultados están acotados entre 7 y 16, es decir que respeta la escala de las flags.

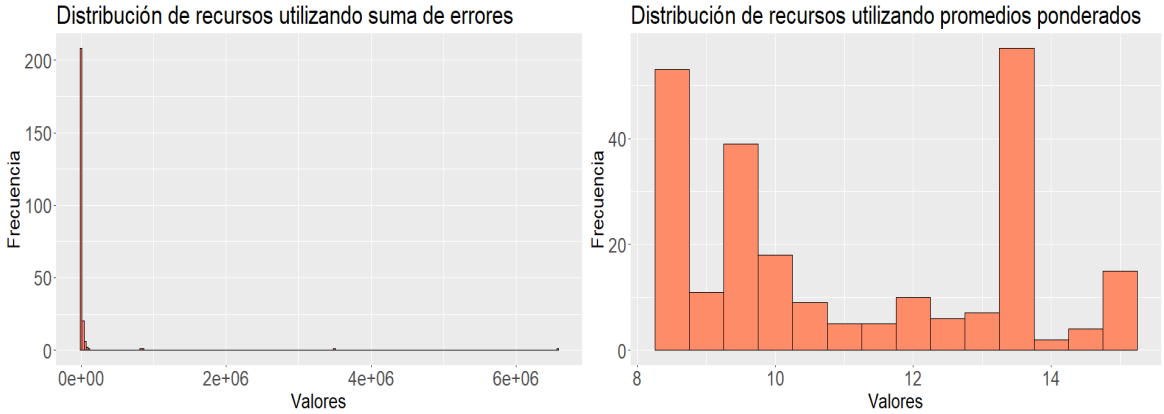

Figura 3. Comparación en la distribución de valores en 320 recursos aleatorios utilizando la metodología del script anterior (izquierda) vs la metodología propuesta durante esta pasantía (derecha).

Este nuevo método se aplicó en el script de tal forma que el resultado tiene la misma estructura que el generado en la pasantía anterior: los recursos están organizados en filas, y en las columnas se muestran las flags y el valor total de su calificación (**Anexo 9.7.7**). En el marco de esta pasantía no fue posible analizar las calificaciones obtenidas bajo este método, ni compararlos con los resultados obtenidos en la pasantía anterior, pero se le entregó al EC-SiB una propuesta de tres pasos, con los que se podría llevar a cabo la comparación.

7. Conclusiones

Los principales resultados de esta pasantía fueron la asignación de autores a los metadatos almacenados en el IPT_Permisos, el diagnóstico de las colecciones científicas en GRSciColl, y las mejoras en el script que se utiliza para revisar la calidad de los recursos ya publicados. Estos logros contribuyeron a los procesos de mejora que el SiB Colombia implementa como parte de sus principales ejes de trabajo. Porque por un lado, se eliminó una barrera técnica que dificultaba el acceso a los recursos asociados a los permisos de recolección; por otro

lado, se aumentó la visibilidad internacional de algunas de las colecciones biológicas del país; y finalmente, se mejoró el mecanismo utilizado para el diagnóstico de calidad de los datos publicados, conforme a los lineamientos del estándar Darwin Core.

En conjunto, dichos resultados también son una contribución a la ciencia abierta de los datos de biodiversidad del país, porque no solo se facilitó el acceso a algunos recursos, sino que también se invirtieron esfuerzos relacionados con garantizar su calidad. Por ejemplo, a partir de las mejoras en el script de revisión de calidad, se estableció un nuevo orden de cuáles son los recursos que deben ser ajustados con prioridad. Por tanto, una vez se ejecute el script y se hagan las revisiones y ajustes respectivos (por próximos pasantes o el EC-SiB), estos recursos contarán con una mejor calidad, mejorando los productos que se deriven de sus uso, tales como toma de decisiones, investigación o educación.

Por otro lado, el diseño e implementación del script que asignó de forma automatizada más de 4300 autores, y los cambios hechos en el script de calidad que reducen su demanda computacional, fueron una optimización a las actividades relacionadas, puesto que permitieron una mayor precisión y eficiencia. Además, al consolidar una documentación clara sobre cada una de las actividades realizadas, se optimiza el próximo proceso de pasantías, pues no se tendrá que destinar mucho tiempo en tratar de entender las cosas que se hicieron, sino que se pueden contextualizar rápidamente y así maximizar el tiempo de ejecución de actividades.

El SiB Colombia, como red nacional de datos abiertos sobre Biodiversidad, se encarga de que los datos que son publicados a través de ellos puedan ser visibles internacionalmente, puedan ser accedidos por la mayor cantidad de audiencias posibles, y cumplan estándares de calidad ampliamente aceptados. Esto, enmarcado en el principio de que no solo se necesita facilitar el acceso a los datos, sino también garantizar que estos sean confiables y puedan ser reutilizados, permite ver que el marco de gobernanza del SiB Colombia está a la altura de iniciativas globales.

Finalmente, este periodo como pasante del EC-SiB, en la línea de Administración de Contenidos enfocada en la Informática de la Biodiversidad, me permitió identificar cuáles son los principales retos que se están enfrentando actualmente alrededor de la publicación y uso de los datos de biodiversidad del país, y que posiblemente se pueden aplicar a otras áreas del conocimiento: saber emplear lenguajes de programación para realizar tareas específicas, comunicarse con diferentes entidades e iniciativas para articular conjuntamente procedimientos que le den mayor visibilidad a la información, y gestionar grandes volúmenes de datos de una forma eficiente y reproducible.

8. Recomendaciones

Durante la actividad del inventario de migraciones, se intentó determinar por qué no se le pudo relacionar de forma automática un correo a 2515 recursos (Hoja *conteos*, **Anexo 9.3**). Se encontró que una gran cantidad de casos se debía a inconsistencias al momento de copiar las razones sociales de las empresas (por ejemplo “SAS”, “S.A.S.”, “S.A.S”; un solo punto de diferencia hace que los programas o procedimientos automatizados los identifiquen como diferentes). Por esto se recomienda buscar medidas para garantizar que estos términos sean escritos de forma homogénea. Una posibilidad es hacer un glosario de términos aceptados, e indicar a cada persona que vaya a crear un nombre en la base de datos del SiB, que siga esa terminología.

Para establecer un mejor flujo de trabajo entre los diferentes grupos de pasantías, se recomienda al EC-SiB que cada pasante documente sus actividades más relevantes. Esto se podría hacer, por ejemplo, mediante una plantilla en la cual se registre el contexto de la actividad, cómo se realizó, cuáles fueron los principales resultados obtenidos, y de forma opcional una explicación de las dificultades encontradas.

Incluso con las modificaciones realizadas sobre el script, las cuales están enfocadas a mejorar la eficiencia del proceso, éste todavía sigue siendo demandante computacionalmente. Una opción que se puede explorar es implementar la ejecución en paralelo del código, lo cual permite utilizar simultáneamente varios núcleos de la computadora. Si bien esto requiere cambios considerables en el script, puede ser de mucha utilidad dada la gran cantidad de recursos que se manipulan. Un texto introductorio para el lenguaje R fue hecho por Radečić (2024). Se pueden buscar alternativas para Python.

Para mejorar aún más la reproducibilidad del script que se utiliza para realizar los diagnósticos de calidad, se recomienda tener una carpeta en la nube con todos los archivos, para que de esta forma, sin importar el computador en el cual se utilice, se pueda utilizar la misma ruta dentro del script y no se tenga que manipular manualmente la ubicación de las carpetas.

Durante la pasantía fue evidente que el lenguaje Python ofrece algunas ventajas con respecto al lenguaje R, especialmente en la gestión de archivos de texto plano y la manipulación de cadenas de texto. Si bien en el curso de programación de la facultad recibí muy buenos fundamentos sobre el pensamiento programativo (lo cual es aplicable a cualquier lenguaje), este solo se enfocó en el lenguaje R. Por estas razones, se le recomienda a la facultad incluir por lo menos una sección para ver los fundamentos de Python, y sus diferencias con el lenguaje R. También se puede incentivar las sesiones extracurriculares enfocadas a la manipulación y análisis de la información de forma programativa.

9. Anexos

9.1 Glosario de términos utilizados

El archivo se puede consultar a través del siguiente enlace: https://docs.google.com/document/d/1EWXJh6blO-m4cYYil8bFc_J6ZBnaCqeLTrefv0YnRis/edit?usp=sharing. Contiene los términos específicos de la temática organizados en orden alfabético.

9.2 GoogleSheets del inventario Migraciones depurado

El archivo se puede consultar a través del siguiente enlace: https://docs.google.com/spreadsheets/d/1d9Fsuva9T0NagMSqUZc65IFcD2IFCmgr-hzzwwdFdBQ/edit?usp=sharing. Está dividido en siete hojas:

- *metadata*: Significado de los colores utilizados en las hojas *Migraciones CR-SiB* y *base_asignación*.
- *Migraciones CR-SiB*: Hoja del inventario filtrada y organizada con los colores.
- *listado*: Listado de publicadores actualizado, comparado con las opciones en IPT_Permisos.
- *base_asignación*: Copia de cuatro columnas de *Migraciones CR-SiB*, considerando solo aquellos recursos sin NA en "Correo organización". También se registra el estado de asignación, algunos comentarios y una comparativa entre el enlace ajustado y el enlace previo.
- *trial_asignación*: Ejemplo de la plantilla utilizada para descargar una selección de filas en formato csv.
- *egAsignacion*: Recursos utilizados para la grabación del video (**Anexo 9.4**).
- *evParrafos*: Tabla del mapeo de 100 recursos aleatorios, con sus enlaces, y la entidad registrada en el párrafo. También se incluye tabla resumen.

9.3 GoogleSheets de las modificaciones realizadas a los titulares

El archivo se puede consultar a través del siguiente enlace: https://docs.google.com/spreadsheets/d/1OxH0ZlF-zlAjsKQ4OvkcctJKkMgtzl3lWLoOgbSnyuQ/edit?usp=sharing. Está dividido en cuatro hojas. En la primera de ellas se da una explicación de las tres hojas restantes y se da la definición de cada una de las columnas utilizadas.

9.4 Carpeta con archivos relacionados al script de asignación

La carpeta se puede consultar a través del siguiente enlace: https://drive.google.com/drive/folders/16iLEyrm2dCOgByp7kWtNf-y37-yKoM28?usp=sharing. Contiene cuatro archivos:

- script_asignacion.py: Script definitivo y organizado. Por confidencialidad con el SiB no tiene la contraseña de las credenciales usadas durante la pasantía.
- explicacion_script.html: Explicación detallada del funcionamiento del script. Se debe descargar para visualizarla adecuadamente.
- ejemplo_script.mp4: Video con una pequeña demostración de la ejecución del script.
- egAsignacion.csv: Archivo de cinco recursos que es utilizado durante la grabación de la demostración.

9.5 GoogleSheets del Inventario de Colecciones depurado

El archivo se puede consultar a través del siguiente enlace: https://docs.google.com/spreadsheets/d/1qPWFG8-t2TG9f0xfnI3fOremw3PID3wwRJRRU91u_V0/edit?usp=sharing. Está dividido en varias hojas, pero la gran mayoría de ellas están ocultas.

- *Revisar*: Hasta la fila 19, muestra el proceso realizado por el Registro Nacional de Colecciones Biológicas (RNC) para crear las colecciones e instituciones que no estaban. De ahí en adelante son los casos especiales que se identificaron al comparar los nombres de las colecciones del inventario con los nombres de las colecciones que están en la hoja *EnlacesCol*
- *PendientesC*: Listado de los recursos que deben ser revisados puntualmente. Se indica la key para los recursos que ya estaban en el inventario, y para todos se documenta (en orden): la cantidad de registros actualizados, la cantidad de registros indexados en GRSciColl, el nombre de la colección, el nombre del recurso, el estado del recurso frente a la indexación en GRSciColl, los dos enlaces respectivos, y una descripción de la situación.
- *Colores*: Explicación de la paleta de colores en la hoja de *Colecciones20221231*, así como una descripción breve de las categorías de la columna "Estado" en la hoja *mapeo RRBB*
- *Colecciones20221231*: Inventario actualizado de las colecciones con corte en diciembre del 2022.
- *Mapeo RRBB*: Inventario de cada uno de los recursos asociados a las colecciones revisadas. No se indica el key de los recursos que no estaban en el inventario, y para todas se coloca (en orden): cantidad de registros actualizados, cantidad de registros mapeados en GRSciColl antes de hacer el ajuste en *CollectionID*, cantidad de registros en GRSciColl después de hacer el ajuste (si hay cambios); Nombre de la colección, título del recurso, y los enlaces respectivos; una columna con las clasificaciones de los 11 casos, que a veces se profundiza en la columna

"Observacion"; finalmente, tres columnas con información del recurso que se obtiene en el IPT.

- *Tot_IPTvsGR*: Tabla que muestra, para cada una de las colecciones encontradas en GRSciColl, el total de registros en el inventario, la cantidad de registros que estaban indexados antes de hacer los ajustes, y la cantidad de registros que quedaron indexados después del ajuste. También se muestran dos columnas que comparan estos valores, y según esto clasifica el estado en que estaba la indexación antes y después del ajuste.
- *RNCvsGR*: Comparativa de los nombres de colección del inventario con respecto a los de GRSciColl, soportado en otras columnas como la organización, key, tipo de colección, nameCollection.
- *EnalcesCol* y *EnlacesIns* tablas con los nombres y enlaces de GRSciColl para las colecciones e instituciones que aparecen para Colombia. Aquellas que tienen valor "Nueva" en la columna "Origen" son aquellas agregadas por el RNC como fruto del diagnóstico.

9.6 Algunos resultados de la pasantía anterior (Jefer Cano)

- Anexo9.6.1: Archivo que contiene el esquema con el cual se asignaron las magnitudes de las Issues & Flags seleccionadas. Archivo llamado "**Anexo 3**" https://docs.google.com/spreadsheets/d/1aX8MwrEYIsC_vK4BKuTHO5agCHs4pCJLO5TiSYZ9LcM/edit?usp=sharing
- Anexo9.6.2: Script desarrollado en la pasantía anterior para calificar los conjuntos de datos con base en la prioridad de los ajustes. Archivo llamado "**ScriptDiagnostico_final_ajustes_RO.py**" https://github.com/JeferDCano/Script-diagnostico-SiB-Colombia/blob/main/ScriptDiagnostico_Final_ajustesRO.py
- Anexo9.6.3. Archivo con los nombres y magnitudes de los Issues & Flags que fueron seleccionados en la pasantía anterior. Archivo llamado "**pesos_issues_final.txt**" https://github.com/JeferDCano/Script-diagnostico-SiB-Colombia/blob/main/pesos_issues_final.txt

9.7 Carpeta con los archivos relacionados al script actualizado

La carpeta se encuentra en el siguiente enlace: https://drive.google.com/drive/folders/1Q2Zoxm15nzEiqYeUP-U4PTjAQthY2XBh?usp=sharing
Contiene los siguientes archivos:

- Anexo9.7.1: Documentación que respalda los cambios hechos en el script.
- Anexo9.7.2: Script con el procedimiento utilizado para comparar Spyder vs reticulate. Debajo de cada prueba los resultados están anotados como comentarios.
- Anexo9.7.3: Script con el procedimiento utilizado para comparar diferentes metodologías para importar archivos con Spyder.

- Anexo9.7.4: Script actualizado, dividido en dos archivos. El principal es “scriptMod”.
- Anexo9.7.5: Tabla basada en el Anexo 9.6.1 con la nueva flag. Archivo llamado “**Tabla con nueva flag**”
- Anexo9.7.6: Archivo csv basado en el Anexo 9.6.3, en el que se inclyó la magnitud de la nueva flag. Archivo llamado “**magnitudes_flags.txt**”.
- Anexo9.7.7: Archivo csv que resultó de la ejecución del script utilizando el nuevo método. Archivo llamado “**dwc_final_completo.csv**”.

10. Bibliografía

Amariles, D., Escobar, D., Gómez Ahumada, M. F., Orrego, Ó., y Soacha-Godoy, K. (2015). Informe *Anual SiB Colombia—2014*. Sistema de Información sobre Biodiversidad de Colombia. http://repository.humboldt.org.co/handle/20.500.11761/34857

Awada, L., Phillips, P. W. B., y Bogdan, A. M. (2022). Governance and stewardship for research data and information sharing: Issues and prospective solutions in the transdisciplinary plant phenotyping and imaging research center network. *Plants, People, Planet*, *4*(1), 84-95. https://doi.org/10.1002/ppp3.10238

Buitrago, L. (2020). *GBIF Issues & Flags*. https://data-blog.gbif.org/post/issues-and-flags/

Canhos, D. A. L., Sousa-Baena, M. S., De Souza, S., Maia, L. C., Stehmann, J. R., Canhos, V. P., De Giovanni, R., Bonacelli, M. B. M., Los, W., y Peterson, A. T. (2015). The Importance of Biodiversity E-infrastructures for Megadiverse Countries. *PLOS Biology*, *13*(7), e1002204. https://doi.org/10.1371/journal.pbio.1002204

Cano, J. D. (2023). Diagnóstico y priorización de datos sobre biodiversidad con problemas de calidad publicados a través del SiB Colombia [Tesis de pregrado inédita]. Universidad de los Llanos.

Díaz, J., Escobar, D., Plata, C., y Ortiz, R. (2022). Informe anual de la Red Nacional del SiB Colombia 2021. Sistema de Información sobre Biodiversidad de Colombia. http://repository.humboldt.org.co/handle/20.500.11761/35906

Díaz, J., Gamboa, J., Buitrago, L., y Escobar, D. (2019). Informe Anual 2018—SiB Colombia. reponame: Repositorio Institucional de Documentación Científica Humboldt. http://repository.humboldt.org.co/handle/20.500.11761/35364

Díaz, J., Plata, C., Ortiz, R., Marentes, E., y Escobar, D. (2023). Informe 2022: Coordinación y participación en la red del SiB Colombia. Sistema de Información sobre Biodiversidad de Colombia. http://repository.humboldt.org.co/handle/20.500.11761/36147

Escobar, D., Buitrago, L., Gamboa, J., y Díaz, J. (2018). Plan estratégico del SiB Colombia. Sistema de Información sobre Biodiversidad de Colombia. http://repository.humboldt.org.co/handle/20.500.11761/35500

Escobar, D., Gómez, N., Soacha-Godoy, K., Grajales, V., y Beltrán, N. (2016). Informe Anual SiB Colombia—2015. Sistema de Información sobre Biodiversidad de Colombia. http://repository.humboldt.org.co/handle/20.500.11761/34856

Faith, D., Collen, B., Ariño, A., Koleff, P., Guinotte, J., Kerr, J., y Chavan, V. (2013). Bridging the biodiversity data gaps: Recommendations to meet users' data needs. Biodiversity Informatics, 8(2), 41-58. https://doi.org/10.17161/bi.v8i2.4126

Gamboa, J., Romero Fernández, J. S., Pino, D., y Barreto, I. (2019). Informe semestral sobre las actividades de mantenimiento y actualización de la infraestructura informática del SiB Colombia. 2019-1. Sistema de Información sobre Biodiversidad de Colombia. http://repository.humboldt.org.co/handle/20.500.11761/35372

GBIF. (2023). GRSciColl. https://scientific-collections.gbif.org/about

GBIF Secretariat. (2022). Introduction to GBIF. https://docs.gbif.org/course-introduction-to-gbif/en/introduction-to-gbif.en.pdf

Gómez-Ahumada, M. F. (2013). Informe Anual SiB Colombia—2012. reponame: Repositorio Institucional de Documentación Científica Humboldt. http://repository.humboldt.org.co/handle/20.500.11761/34860

La Salle, J., Williams, K. J., y Moritz, C. (2016). Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1702), 20150337. https://doi.org/10.1098/rstb.2015.0337

Marentes, E., Ortiz, R., Lozano, J., y Plata, C. (2023). Cómo cargar los datos al IPT, Ciclo de formación. https://biodiversidad.co/formacion/laboratorios/carga-datos-ipt

Ortiz, R., Marentes, E., y Leuro, N. (2024). Reporte mensual SiB Colombia—Enero 2024. reponame: Repositorio Institucional de Documentación Científica Humboldt. http://repository.humboldt.org.co/handle/20.500.11761/36282

Plata, C., Ortiz, R., Marentes, E., y Díaz, J. (2022). Informe anual de conjuntos de datos e información publicados o actualizados a través del SiB Colombia 2021. Sistema de Información sobre Biodiversidad de Colombia. http://repository.humboldt.org.co/handle/20.500.11761/35910

Ministerio del Medio Ambiente. (1994). Decreto 1603 de 1994. Por el cual se organizan y establecen los Institutos de Investigación de Recursos Biológicos «Alexander von Humboldt», el Instituto Amazónico de Investigaciones «SINCHI» y el Instituto de Investigaciones Ambientales del Pacífico «John von Neumann».

Ministerio de Ambiente y Desarrollo Sostenible. (2013a). Decreto 1376 de 2013. Por el cual se reglamenta el permiso de recolección de especímenes de especies silvestres de la diversidad biológica con fines de investigación científica no comercial.

Ministerio de Ambiente y Desarrollo Sostenible. (2013b). Decreto 3016 de 2013. Por el cual se reglamenta el Permiso de Estudio para la recolección de especímenes de especies silvestres de la diversidad biológica con fines de Elaboración de Estudios Ambientales.

Radečić, D. (2024). R doParallel: A Brain-Friendly Introduction to Parallelism in R. https://www.appsilon.com/post/r-doparallel

Salinas, P. Y. (2022). Apoyo en el proceso de limpieza retrospectiva de datos publicados a través del SiB Colombia para mejorar su visibilidad, consulta y uso [Tesis de pregrado, Universidad Distrital Francisco José de Caldas]. http://repository.udistrital.edu.co/handle/11349/30298

SiB Colombia. (2020a). Guía para publicar datos e información. https://biodiversidad.co/compartir/guia-para-publicar/

SiB Colombia (Director). (2020b, septiembre 16). Capítulo 2 ¿Qué datos sobre nuestra biodiversidad se pueden publicar a través del SiB Colombia? https://www.youtube.com/watch?v=_f4gGfIBN3U

SiB Colombia. (2021a, diciembre 29). ¿Qué es el SiB Colombia? https://biodiversidad.co/acercade/sib-colombia/

SiB Colombia. (2021b, diciembre 25). Crear Compartir Transformar. https://biodiversidad.co/recursos/acceso-abierto/

SiB Colombia. (2022a). Programa de pasantías. https://biodiversidad.co/comunidad/formacion/programa-pasantias/

SiB Colombia. (2022b). Tipos de datos e información. https://biodiversidad.co/compartir/tipos-de-datos/

Soberón, J., y Peterson, T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1444), 689-698. https://doi.org/10.1098/rstb.2003.1439

Vicente-Saez, R., y Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. Journal of Business Research, 88, 428-436. https://doi.org/10.1016/j.jbusres.2017.12.043

Villa, C. M., Morales, D. P., y Rozo, S. M. (Eds.). (2023). Informe de Gestión Institucional 2022. Instituto de Investigación de Recursos Biológicos Alexander von Humboldt. http://repository.humboldt.org.co/handle/20.500.11761/36194

Wheeler, Q. D., Raven, P. H., y Wilson, E. O. (2004). Taxonomy: Impediment or Expedient? Science, 303(5656), 285-285. https://doi.org/10.1126/science.303.5656.285

Wieczorek, J. (2023). Releases · tdwg/dwc. GitHub. https://github.com/tdwg/dwc/releases