Memorias del Taller CESP 2017 / SiB Colombia - VertNet

Compartiendo experiencias y herramientas en calidad de datos sobre biodiversidad (11-15 de septiembre)

Participantes (toda la semana *)

- **John Wieczorek** - VertNet (Arquitecto de Información - Museum of Vertebrate Zoology, UC Berkeley y Museum of Comparative Zoology, Harvard University.) *
- **Paula Zermoglio** - VertNet (Investigadora Asociada -Instituto de Ecología, Genética y Evolución de Buenos Aires (IEGEBA-CONICET), Universidad de Buenos Aires, Argentina y Université François Rabelais de Tours, Francia)*
- **Leonardo Buitrago** - SiB Colombia (Administración de contenidos) *
- **Victoria Arciniegas** - SiB Colombia (Cooperación) *
- **Oscar Duque** - SiB Colombia (Tecnologías de la Información) *
- **Ricardo Bastidas** - Instituto Humboldt (Tecnologías de la Información) *
- **Iván González** - Instituto Humboldt (Indicadores de Biodiversidad) *
- **Ricardo Ortíz** - Investigador, Especialista en SIG y georreferenciación *
- **Dairo Escobar** - SiB Colombia (Coordinador)
- **Javier Gamboa** - SiB Colombia (Reporte y Síntesis)
- **Carolina Castro** - Instituto Humboldt (Infraestructura y calidad de datos - PEM)*
- **Carlos DoNascimiento** - Instituto Humboldt (Curador Colección de Peces de agua dulce)
- **Kevin Borja** - Instituto Humboldt (Tecnologías de la Información / Colecciones Biológicas)

Propósitos

Principal

Transferencia de la experiencia de VertNet sobre la mejora automática de la calidad de los datos de biodiversidad en los flujos de publicación del SiB Colombia (SiB) y el Instituto Humboldt (IAvH).

Específicos

- Conocer e implementar la herramienta "Darwin Core Data Migrator Toolkit" dentro de los procesos de calidad de datos en el SiB y el IAvH.
- Realizar un intercambio de experiencias entre VertNet, el SiB Colombia y el IAvH sobre procesos internos y herramientas de calidad de datos que permitan optimizar y mejorar la información sobre biodiversidad que se publica.
- Generar la documentación necesaria de la herramienta "Darwin Core Data Migrator Toolkit" y la traducción de la misma al español, para facilitar el uso de la herramienta a los participantes y las instituciones publicadoras de datos de la comunidad de habla hispana, haciéndolo extensivo a los nodos GBIF (en español).
- Realizar un intercambio de experiencias sobre el proceso de digitalización en las colecciones biológicas y cómo mejorar sus flujos de trabajo.

Agenda general

HORARIO	LUNES 11	MARTES 12	MIÉRCOLES 13	JUEVES 14	VIERNES 15
8:00 – 9:00	Bienvenida y recorrido Venado de Oro	Puesta en marcha Data *Migrator Toolkit*: Ejemplo	Sesión de trabajo *Data Migrator Toolkit*	Procesos de documentación *Data Migrator Toolkit*	Flujo de trabajo colecciones biológicas / I2D
9:00-10:00	Revisión de agenda. Flujo de trabajo y contexto SiB Colombia				Intercambio de experiencias proceso de digitalización en las colecciones
10:00-10:30	Receso	Receso	Receso	Receso	Receso
10:30-11:30	Flujo de trabajo y contexto VertNet	Puesta en marcha *Data Migrator Toolkit:* Ejemplo (continuación)	Sesión de trabajo *Data Migrator Toolkit* (continuación)	Intercambio de experiencias sobre calidad de datos	Perspectivas de mejoramiento en los flujos de trabajo - colecciones biológicas / I2D
11:30-12:30	Introducción y requerimientos de instalación *Data Migrator Toolkit*				Conclusiones y cierre flujo de trabajo colecciones biológicas / I2D
12:30-14:00	Almuerzo	Almuerzo	Almuerzo	Almuerzo	Almuerzo
14:00-15:00	Instalación *Data Migrator Toolkit*	→ Traslado a Universidad Javeriana para la charla	Mantenimiento de vocabularios *Data Migrator Toolkit*	Perspectivas de mejoramiento en los flujos de trabajo - calidad de datos: - Kurator - Mantenimiento de Vocabularios	
15:00-16:00	Revisión general de los componentes del *Data Migrator Toolkit*	Charla: *"El bueno, el malo y el no tan lindo. ¿Cómo lidiar con datos de biodiversidad?"*	Retroalimentación *Data Migrator Toolkit*	Evaluación y cierre de la herramienta *Data Migrator Toolkit* y calidad de datos	
16:00-17:00		Paula y John			

Memorias

Aspectos generales del taller

Durante la reunión se trabajó utilizando distintos documentos comunes alojados en una carpeta compartida en Google Drive: CARPETA CESP-SIB. Dicha carpeta contiene documentos generales, tales como la agenda de la reunión, y distintas subcarpetas en las que se organizaron archivos de documentación (lista y en proceso), presentaciones, etc.

Los recursos utilizados relacionados con el migrador se encuentran en los siguientes links:

- Migrator, repositorio GitHub: https://github.com/VertNet/toolkit
- Vocabularios, repositorio GitHub: https://github.com/tucotuco/DwCVocabs

Además, durante el taller se acordó la creación de un repositorio en GitHub como el canal de comunicación entre el SiB Colombia, Instituto Humboldt y VertNet para consolidar el proceso de implementación y puesta en marcha de la herramienta, documentando novedades y posibles fallos o mejoras (issues) a lo largo del proceso: CESP-GBIF-SiB-VertNet.[1]

A partir de las actividades desarrolladas en el taller y de lo acordado para llevar a cabo en los meses subsiguientes, se estableció un cronograma tentativo para la culminación de algunas actividades.

Actividades desarrolladas y resultados obtenidos durante el taller

Lunes 11

1) Se presentaron los **flujos de trabajo y contextos** del SiB Colombia y de VertNet. Se discutieron similitudes y diferencias de abordajes utilizados. Presentaciones: VertNet, SIB Colombia.

2) Se presentó la **herramienta Migrador** y sus componentes principales en una sesión *demo*.

3) Se **instaló el Migrador** en las computadoras de los asistentes. Para ejecutar los migradores se requieren cierta configuración regional y del programa (Access), y tener instaladas utilidades Unix. Se configuraron las computadoras de trabajo y se documentaron los pasos a seguir para llevar a cabo dicha configuración para utilizar el Migrador. Los pasos y las explicaciones correspondientes se capturaron en castellano y están disponibles en el documento Pasos Versión ES, en la sección "I. Configurando la computadora para hacer el trabajo". Dicho documento es la base para la traducción al

[1] En este repositorio ya se están capturando *issues* con las actividades a realizar y los problemas encontrados en la ejecución del migrador.

castellano de la guía de uso paso a paso del Migrador. [2] La versión en inglés se encuentra disponible aquí Pasos Versión EN.
Para comprobar que las computadoras pudieran ejecutar el Migrador, se trabajó con un migrador ya listo, provisto por VertNet y basado en un conjunto de datos reales.

Martes 12

1) Se inició la **práctica de uso del Migrador** utilizando un conjunto de datos ejemplo provisto por el SIB Colombia. Dicho conjunto de datos fue de estructura sencilla, tal de demostrar las funciones básicas del migrador.

2) Se llevó a cabo la **charla**: "El bueno, el malo y el no tan lindo. ¿Cómo lidiar con datos de biodiversidad?", en la Universidad Javeriana, con asistencia de alrededor de 30 personas.

Miércoles 13

1) Se continuó con la **práctica de uso del Migrador**. La práctica se extendió durante la mayor parte de la jornada. A partir de esta práctica se plantearon diferentes modificaciones necesarias al Migrador. Algunas de dichas modificaciones fueron incorporadas durante el transcurso del taller, mientras que otras fueron señaladas para ser realizadas en las semanas siguientes. Una de las modificaciones ya incorporadas consiste en la reducción del número de consultas que deben personalizarse dado un reajuste de los macros utilizados por el migrador.

2) **Vocabularios**: el tema fue abordado durante las jornadas anteriores y posteriores de manera saltatoria y recurrente. Se discutieron distintas alternativas de manejo de vocabularios, entre las cuales se considera cruzar los vocabularios en uso por el SIB y el Instituto Humboldt con los utilizados por VertNet para crear un repositorio único. Se acordó que se realizarían cambios al proceso de fusión de vocabularios del migrador ("Merger") y se documentaría correspondientemente[3]. Se decidió que el tema vocabularios sería abordado más adelante en un contexto más general. En particular, se estipuló que se llevaría la discusión del tema a la reunión de TDWG de este año en Ottawa (Oct 2017), durante la cual están planeadas varias instancias de discusión de vocabularios.[4] Se planea una reunión del grupo CESP SIB-VertNet post-TDWG para discutir los pasos a seguir referidos a la confección y manejo de vocabularios. Mientras tanto, se acordó que se llevaría a cabo una evaluación del estado actual de los

[2] Durante el transcurso del taller, por fuera del horario del mismo, se llevó a cabo la traducción de esta parte del documento a la versión en inglés de la guía de uso paso a paso del Migrador, incorporando así nuevo material a la documentación en su idioma original. Guía completa EN. Las guías en inglés y castellano están actualmente en revisión, y serán incorporadas en el repositorio del Migrador en GitHub tan pronto como estén listas.
[3] Estas modificaciones y la correspondiente documentación están actualmente en proceso.
[4] Durante esta misma reunión se llevará a cabo la reunión del Data Quality Interest Group, dentro del cual se planea presentar un Task Group especialmente para tratar la construcción de vocabularios controlados, liderado por P. Zermoglio.

vocabularios mantenidos por VertNet en cuanto a potenciales necesidades de traducción al castellano de los términos estándar.[5]

3) Se discutieron los **procesos de documentación** de la herramienta, identificando a) documentos ya existentes y completos en inglés y que deben ser traducidos, b) documentos en inglés que deben ser revisados para luego ser traducidos, y c) documentación nueva que debe ser generada a partir de este proyecto en ambos idiomas.
    a) ***Documentos existentes***
        i) ***Reportes***: Dentro de los migradores existen reportes que capturan los resultados del proceso y los cambios introducidos. Entre los archivos del migrador existe un documento que explica la estructura de los reportes. Se acordó durante el taller adicionar una versión en castellano de la descripción de los reportes, como parte estructural del archivo de migradores, para su uso posterior por comunidades no angloparlantes. La traducción de dicho documento fue iniciada previamente al desarrollo de este taller. Durante el taller se llevó a cabo una primera revisión de la traducción en un documento compartido.[6]
        ii) ***Migradores - documento descriptivo general para usuarios***. La versión original en inglés de este documento, disponible en GitHub, fue revisada (Versión EN). Durante el taller se completó la traducción del documento Versión ES. Además, durante el taller pero por fuera del horario del mismo se generaron gráficos explicativos de los procesos llevados a cabo por el migrador en versión castellano (versión gráfica ES).[7]
        iii) ***Migradores - documento descriptivo detallado de los pasos*** a seguir para correr el migrador (contiene la parte de configuración de la computadoras y de ejecución del migrador). Al momento de finalizado el taller, se contaba con dos versiones, una completa en inglés y otra versión traducida al castellano sólo en parte. Pasos versión EN; Pasos versión ES. Se acordó revisar la versión en inglés, dado que se incorporarían más cambios al migrador, y que posteriormente se introducirían dichos cambios a la versión en castellano para completar la traducción del documento.
    b) ***Documentación nueva:*** Se acordó durante el taller generar la siguiente documentación en el lapso de los próximos meses:
        i) Documentación referida específicamente a ***qué hace cada tipo de consulta/tabla*** dentro del migrador (e.g., legacies).
        ii) ***Conjunto de datos de prueba*** (datos ficticios, que cubran los tipos de errores que el migrador detecta y corrige) y un migrador ya ejecutado sobre ese conjunto de datos de prueba, contra el cual poder testear a

[5] Al día de la fecha (Oct 2017) se ha realizado una evaluación de los vocabularios en cuanto al número de valores por término que podrían potencialmente ser traducidos al castellano (disponible en: VN Vocabs Translation Scope). Se discutirán los pasos a seguir a este respecto en la siguiente reunión online.
[6] Luego de la finalización del taller se llevaron a cabo modificaciones al Migrador que incluyeron cambios en los reportes generados. Dichos cambios fueron incorporados a las versiones en inglés y castellano de explicación de los reportes y se produjeron documentos pdf que ya se encuentran disponibles en GitHub: versión EN, versión ES.
[7] Luego de la finalización del taller se llevó a cabo la traducción de la representación gráfica del Migrador, que ya se encuentra disponible en el Wiki del repositorio GitHub de la herramienta.

manera de práctica el armado de migradores. Dicha documentación será adicionada a la documentación general sobre la herramienta.

iii) ***Documento para el público en general***, explicativo sobre los Migradores y el mejoramiento de la calidad de datos en general, que muestre el valor agregado de la utilización de la herramienta.

iv) ***Manejo de vocabularios***. Documentación referida a la herramienta "Merger". Dado que esto estará sujeto a cambios en el Merger, la producción de la documentación pertinente está supeditada a la estabilización del Merger como proceso. Además, se considera que discusiones más amplias respecto a la construcción y manejo de vocabularios puedan ser factores de retraso en la generación de esta documentación.

v) ***Ejemplos gráficos para cada tipo de datos.*** Dados los esquemas generales del funcionamiento del migrador que se produjeron durante el taller (versión en ES), se propuso generar gráficas similares que demuestren el proceso específicamente para ciertos tipos de datos (e.g., fechas, taxonomía, geografía, etc.).

Jueves 14

1) Se realizó un intercambio de **experiencias sobre calidad de datos**, durante el cual se mostraron flujos de trabajo para mejorar la calidad de los datos y distintas herramientas creadas por miembros del equipo, tanto del SIB Colombia como del Instituto Humboldt. Entre ellas se presentó un evaluador de calidad de datos capaz de detectar errores en conjuntos de datos basado en Google Spreadsheets y una herramienta de modelado de distribución de especies con chequeo taxonómico y geográfico escrita en R, además de protocolos de georreferenciación de localidades. El detalle de las herramientas y procesos vistos se presenta en la Tabla 1.

Tabla 1. Intercambio de experiencias sobre calidad de datos: procesos y herramientas presentadas durante el taller.

Herramienta o recurso	Organización encargada	Presentador
Documentos SiB	SiB Colombia	Leonardo
Plantillas de publicación SiB	SiB Colombia	Leonardo
Wiki de publicación	SiB Colombia	Leonardo
Guías Calidad de datos	SiB Colombia	Leonardo
T-Rex	SiB Colombia	Leonardo
Validaciones y vocabularios controlados	Instituto Humboldt	Carolina
Validador Calidad de datos	SiB Colombia	Oscar
Geo-validador	SiB Colombia	Leonardo
Protocolo Georeferenciación	SiB Colombia/ Instituto Humboldt	Ricardo
Herramienta Geo IAvH	Instituto Humboldt	Iván
Biomodelos	Instituto Humboldt	Iván

2) Se presentó la **herramienta Kurator** en el contexto de las perspectivas de mejoramiento de flujos de trabajo y calidad de los datos. Presentación Kurator.

3) Se llevó a cabo la **evaluación del taller y el cierre** de la sección correspondiente al Migrador y a calidad de datos. Se identificaron ventajas y desventajas relacionadas con la herramienta, que se presentan en la Tabla 2.

Tabla 2. Evaluación de la herramienta Migrador.

VENTAJAS	DESVENTAJAS	SOLUCIÓN
Completa. Validación de calidad de datos en todo el espectro DwC.	Compleja.	1. Simplificar tareas dentro del migrador. 2. Entrenamiento.
Permite estructurar los datos originales con todos lo elementos DwC.	Estructuración requiere un nivel de detalle y cuidado importante.	1. Tener conocimiento de los datos fuente y de los términos del estándar DwC.
Se pueden emplear secciones para mejoramiento de calidad de datos particular.	NO es una herramienta Universal (e.g., no incluye georreferenciación, actualmente más apta para angloparlantes).	1. Mejorar la documentación. 2. Traducir a otros idiomas. 3. Incorporar vocabularios en otros idiomas. 4. Utilizar en combinación con otras herramientas.
Es una herramienta diseñada para los administradores, dada su complejidad.	NO es una herramienta diseñada para los publicadores, dada su complejidad.	1. Aplicar a nivel agregador. 2. Utilizar herramientas de limpieza de datos previas (a nivel proveedor) más sencillas. 3. Entrenamiento.

Viernes 15

1) Se presentaron los **flujos de trabajo de las colecciones biológicas** / I2D del instituto Humboldt.

2) Se realizó un **intercambio de experiencias en cuanto al proceso de digitalización** de colecciones biológicas y a cómo optimizar el rendimiento. Se discutieron diferentes alternativas y se compartieron recursos (e.g., reporte JRS sobre proyectos para compartir datos de biodiversidad de la fundación JRS; digitalización de datos vía crowdsourcing: Notes from Nature; presentación con estimación de tasas de digitalización; etc.)

Consideraciones finales

MIGRADOR
Se presentó y realizó la instalación y puesta en marcha de la herramienta Data Migrator toolkit. Durante el uso e implementación de la misma se establecieron modificaciones necesarias para implementar sobre la herramienta, muchas de ellas realizadas dentro de esta semana de trabajo y otras a realizarse posterior a la misma de acuerdo a un plan de pruebas y documentación en los repositorios de GitHub que se establecieron en el taller.

VOCABULARIOS CONTROLADOS
El tema de la implementación y uso de vocabularios controlados en español dentro de la herramienta e integrados con los ya establecidos en inglés y otros idiomas resulta vital para las validaciones que realiza el migrador y es un aspecto a abordar y solucionar durante todo el periodo de implementación y cierre del proyecto.

DOCUMENTACIÓN
Se compiló y tradujo gran parte de la documentación de la herramienta en español previo al taller y durante su desarrollo, esto permite no solo ir consolidando la información del migrador en habla hispana sino que ayudará a mejorar la documentación original de la herramienta en inglés.

INTERCAMBIO DE EXPERIENCIAS
El intercambio de experiencias en calidad de datos que se dio durante la semana de trabajo permitió ver posibles sinergias e implementación de módulos o ideas que dan la opción de contar con un migrador más robusto o la integración o mejora de herramientas ya desarrolladas en calidad de datos.