Herramienta de Migración de Datos de VertNet: Data Migrator Toolkit

Introducción

Los Migradores de Datos son bases de datos construidas en Microsoft Access y personalizadas para enlazar o importar datos fuente. Los Migradores:

- procesan datos y crean un archivo CSV Darwin Core y opcionalmente un archivo CSV de Extensión de Medios (Media Extension),
- agregan vocabularios nuevos a un Gestor de Vocabularios,
- resuelven vocabularios revisados a valores estándar presentes en el Gestor de Vocabularios,
- remueven caracteres no visibles (non-printing characters) problemáticos,
- llevan un registro de los cambios hechos a los datos fuente, y,
- crean reportes que detallan problemas potenciales y recomiendan cambios para hacer en la fuente de datos original.

Plantillas del Migrador

El siguiente repositorio en Github contiene la última versión de la plantilla del migrador: https://github.com/VertNet/toolkit
Para determinar la versión de la plantilla se puede consultar la fecha de la entrada más reciete en el archivo ChangeLog.txt (https://raw.githubusercontent.com/VertNet/toolkit/develop/ChangeLog.txt). Cada cambio que se realiza en la plantilla es registrada en este archivo, de modo que sea sencillo actualizarse a versiones más nuevas del migrador mirando los cambios que han sido incorporados en la plantilla desde la creación de la primera versión del migrador.
La plantilla consiste en bases de datos de Access, scripts y carpetas de archivos utilizados para procesar los datos fuente y obtener archivos listos para transferir a un recurso en una instancia del Integrated Publishing Toolkit de GBIF (IPT). La plantilla es la base de cada migrador y es personalizada para cada recurso distinto en el IPT. El migrador puede combinar múltiples conjuntos de datos en un conjunto de datos agregado para poder crear un recurso único en el IPT.

Personalización del Migrador

Cada fuente de datos es única y requiere algún tratamiento de personalización para transformarla a Simple Darwin Core (y a una extensión opcional de medios -Media extension, si es necesario) y para prepararla para su carga como un recurso en IPT. La mayoría de la personalización ocurre en consultas y macros dentro de dos bases de datos de Access dentro del migrados, contenidas dentro de la carpeta “templates”.

Versión en castellano preparada por SIB Colombia y VertNet bajo un proyecto GBIF-CESP 2017.

La primera de estas bases de datos se llama “DwC2ExtractTemplate-XXXX.mdb”, donde XXXX es uno de “Audio”, "Aves", "Eggs", “Ent”, "Fish", "Fossils", “Fungi”, "Herps", “Inverts”, "Mammals", “Plants”, "Verts" (dependiendo de si queremos que el migrador actúe específicamente sobre alguno de estos grupos). Esta base de datos es utilizada para enlazar la fuente de datos original al migrador y para llevar a cabo los pasos preliminares necesarios para transformar los datos originales en campos de Darwin Core.

La segunda base de datos, llamada AggregatorTemplate.mdb, contiene consultas y un macro ("Aggregate and Export") para combinar distintas fuentes de datos y crear un archivo CSV Simple Darwin Core agregado, listo para subir a un recurso en IPT. El Agregador debe ser invocado para crear el archivo CSV Darwin Core incluso si se trabaja con una única fuente de datos. El macro debe ser modificado para incluir las fuentes de datos migradas relevantes (cualquier combinación de “Audio”, "Aves", "Eggs", “Ent”, "Fish", "Fossils", “Fungi”, "Herps", “Inverts”, "Mammals", “Plants”, "Verts").

La descripción de los pasos a seguir para personalizar un migrador para un conjunto de datos nuevo y de cómo ejecutar cada paso se proveen en el archivo "README_Instrucciones de uso_ES.pdf" (in https://github.com/VertNet/toolkit).

Vocabularios

El Migrador enlaza a y utiliza una base de datos de Vocabularios (VocabulariesMaster.mdb), que puede encontrarse en: https://github.com/tucotuco/DwCVocabs/tree/master/master. La base de datos de Vocabularios contiene tablas de búsqueda par varios términos Darwin Core, individuales y combinados, y provee valores estandarizados para valores originales encontrados en las fuentes de datos. Las tablas de búsqueda del Vocabulario son pobladas cada vez que se ejecuta un migrador, adicionando a las tablas de búsqueda de los vocabularios los valores que nunca se habían encontrado antes. Estos nuevos valores deben ser revisados por alguien que tenga derechos de acceso de administrador sobre los vocabularios y valores estandarizados para cada nuevo término.

Los vocabularios son consultados durante el curso del proceso de migración de datos para reemplazar valores o términos no estandarizados por sus equivalentes estandarizados. Para asegurarse que los valores estandarizados son incluidos, el migrador debe ser corrido una vez para poblar la base de datos de Vocabularios con aquellos valores nunca antes vistos, y luego nuevamente, una vez que se han resuelto todos esos valores, para llevar a cabo las sustituciones por los valores estandarizados.

Para resolver los vocabularios, existe una base de datos en Access separada, VocabulariesManager.mdb. Esta base de datos enlaza a las tablas en VocabulariesMaster.mdb y tiene una serie de consultas y macros para facilitar el manejo de los vocabularios.

Los contenidos más actualizados de las tablas de búsqueda de los Vocabularios son exportados como archivos CSV y mantenidos en el repositorio en GitHub DwCVocabs: https://github.com/tucotuco/DwCVocabs/tree/master/vocabs.
Estos archivos son actualizados en GitHub luego de cada nueva resolución de vocabularios, usualmente luego de la creación de un migrador para una nueva fuente de datos.

Revisión

Los migradores generan un archivo CSV Simple Darwin Core (y opcionalmente un archivo CSV de Extensión de Medios -Media Extension), y se considera mejores prácticas compartir este archivo con el proveedor de datos para que éste los revise antes de que sean cargados al IPT y se tornen públicos. Esto permite al publicador de datos comprender cómo se verán los datos antes de que autoricen su publicación.

Durante el procesamiento de datos del migrador se generan una serie de reportes para cada fuente de datos, que son acumulados en una carpeta de reportes. Estos reportes muestran dónde se encontraron los problemas potenciales con la calidad de los datos, el formato, o la estandarización. Se considera mejores prácticas compartir estos reportes con el proveedor de los datos antes de la primera publicación del conjunto de datos en el IPT, para que el proveedor pueda determinar si quiere realizar cambios en la fuente basado en estos reportes antes de publicar los datos. Consecuentemente, los datos son publicados una vez que los datos originales se consideran aptos para la publicación, a veces luego de múltiples ciclos de ejecución del migrador y revisión.

Entre los reportes compartidos con los proveedores de datos están aquellos que informan:

- números de catálogo duplicados o faltantes
- años, meses y días fuera de rango
- nombres de regiones geográficas no estandarizados y geografía indeterminada
- nombres taxonómicos no estandarizados al rango de género y superior
- cambios realizados en los datos publicados vs. el verbatim original
- caracteres no imprimibles (non printing characters) en el contenido de los datos que comprometen la integridad de los datos cuando son formateados para compartir archivos CSV.

Una descripción detallada de los reportes puede encontrarse dentro del migrador, en la carpeta “reports”, bajo el nombre “Explicación de Reportes”.

Mantenimiento del Migrador

Hasta ahora, VertNet ha mantenido un archivo para todos y cada una de las fuentes de datos de las organizaciones que optan por el servicio de migradores. Las innovaciones son desarrolladas con la personalización de cada nuevo migrador. Estas innovaciones son capturadas en la plantilla (y en el repositorio en GitHub) y son registradas en el archivo ChangeLog.txt file (https://raw.githubusercontent.com/VertNet/toolkit/develop/ChangeLog.txt).

VertNet ha mantenido también archivos de vocabularios. Actualmente, a medida que más personas utilizan el migrador, estamos explorando nuevas maneras para hacer de esos vocabularios un esfuerzo contributivo, aprovechando el beneficio de la resolución de vocabularios por parte de todos los participantes y encontrando nuevas maneras de integrar los valores estandarizados para que todos podamos utilizarlos.