AUTOMATIZANDO LA CARGA DE DATOS EN EL PORTAL SIEMAC

Sergio Delgado Quintero
Instituto Canario de Estadística, ISTAC
sdelquin@gmail.com

Samuel Baute García
Arte Consultores Tecnológicos S.L, ARTE
sbaugar@arte-consultores.com

Rita Díaz Adán
Arte Consultores Tecnológicos S.L, ARTE
rdiaada@arte-consultores.com

PALABRAS CLAVE
Python, Automatización, ETL, Wordpress, Scraping

RESUMEN

sie-mac.org es el portal de estadísticas de la Macaronesia (Azores, Canarias, Cabo Verde, Madeira e Islas Salvajes). Se trata de una página gestionada desde el ISTAC y que tiene como objetivo difundir los datos estadísticos de esta región agrupados por temáticas y mediante un fácil acceso.

La principal fuente de datos de esta web es Eurostat la cual ofrece información relevante sobre gran variedad de temáticas vinculadas con regiones de Europa. En una fase inicial del proyecto estos datos eran cargados manualmente en la plataforma Wordpress que soporta la página sie-mac.org. Sin embargo, esto generaba una problemática asociada: seguimiento de cambios y nuevos datos en la fuente, gasto de tiempo en actualizar los datos, desfase en la sincronización de la información, tratamiento manual de los datos de origen, etc.

Es por ello que surge la necesidad de automatizar todos estos procesos y, desde esa premisa, nace siemactk (https://github.com/eDatos/siemactk). Este servicio se encarga de hacer un scraping a la web de Eurostat para descargar todos los datasets de interés vinculados con la Macaronesia y agrupados por temáticas. Una vez que los datos se han descargado, se procede a un filtrado, limpieza y recodificación de los mismos. A continuación, se generan ficheros de salida tanto en formato json como en formato tsv, que, a su vez, tienen versiones en castellano y en portugués. Estos ficheros de salida se suben a un bucket de Google Cloud Storage desde donde podrán ser consumidos por el plugin wpDataTables de Wordpress para rellenar las tablas que se presentarán en la web SIEMAC. Y todo ello haciendo uso de una hoja de cálculo Google Sheet en la que se especifican tanto las recodificaciones como las rutas de los ficheros a descargar desde Eurostat. El proceso incluye al final una notificación por correo electrónico en la que se

indican las URLs públicas de destino de todos los datasets ya procesados. En su totalidad estamos hablando de un proceso que no llega a los 10 minutos de ejecución.

El servicio se puede lanzar en cualquier momento, pero actualmente está programado para ejecutarse a primero de mes. Esto garantiza que los datos reflejan la realidad de la información presente en Eurostat y libera totalmente del trabajo manual y seguimiento de cambios que se hacía en la fase inicial del proyecto. Además, al ser un proceso automático, se reducen considerablemente los errores humanos de transcripción, filtrado y procesado de la información.