Islas Canarias
Del 15 al 19 de noviembre de 2021

# Automatizando la carga de datos en el portal SIEMAC

**Rita Díaz Adán**
Arte Consultores
rdiaada@arte-consultores.com

**Sergio Delgado Quintero**
ISTAC
sdelquin@gmail.com

**Samuel Baute García**
Arte Consultores
samuelbaute@gmail.com

## Introducción

En la presente ponencia se describe el sistema creado por el ISTAC y Arte para automatizar datos procedentes del Eurostat y servirlos normalizados, categorizados y actualizados en el portal sie-mac.org.

El portal de datos sie-mac.org nace a raíz de un proyecto INTRREG III-B del mismo nombre y donde uno de los objetivos era la creación de un portal web a través del cual se pudiese publicar información estadística conjunta de la Macaronesia a la vez que específica de cada Archipiélago.

Este portal lleva en funcionamiento desde el año 2006 y desde la creación del mismo ha sido gestionado por el ISTAC. Esta gestión incluía tanto la gestión tecnológica como la de carga y gestión de los datos estadísticos existentes. Esta última parte conllevaba una gran carga de trabajo para los técnicos del ISTAC.

Con el paso de los años, tanto el sistema como la plataforma web exigían una modernización que facilitara el mantenimiento de la página web y la actualización de los datos, ya que ambos habían quedado obsoletos.

Debido a la asimetría de sistemas que se había generado con el paso de los años en los distintos archipiélagos y en sus sistemas de información, la tarea de unirlos de nuevo bajo el mismo paraguas se vislumbraba como un proyecto de largo recorrido e igual de dilatado gasto.

Buscando evitar transitar ese camino, se busca una alternativa a la manera “evolutiva” de abordar el proyecto, buscando un nuevo camino en el que obtener el mismo resultado, tratando de apoyarse todo lo posible en automatizaciones y herramientas de trabajo más actuales que minimizaran el esfuerzo humano.

## Objetivos

Con el nuevo proyecto para el portal sie-mac se requería una modernización completa del sistema de forma que se consiguiera tener el portal actualizado, con datos de calidad y actualizados puntualmente pero sin que esto supusiera un esfuerzo añadido para el personal del ISTAC.

Así mismo, se quería también ampliar la cantidad de datos que se publicaban. Hasta ese momento existían un conjunto de series armonizadas entre las diferentes oficinas estadísticas pero era un subconjunto de la información estadística que podía ser relevante publicar de forma conjunta.

Por otra parte, se buscaba no sólo actualizar los contenidos del portal sino también la infraestructura del mismo. En este sentido se necesitaba disponer de un sitio web fácil de editar y mantener, que permitiese la inclusión de información estática y que diese la posibilidad de visualizar los datos estadísticos.

## Metodología

A la hora de comenzar el proyecto, la primera decisión a tomar fue la nueva plataforma que haría de base para la gestión del contenido de la web, tanto el contenido estático (información presente en la web previa de SIE-MAC, noticias en formato blog, etc.) como los datos, que se actualizarían con regularidad.

El anterior gestor de contenidos había sido realizado a medida y rescatarlo u optimizarlo era una tarea que no ofrecía ninguna ventaja.

Por las amplias posibilidades, extendido uso y por contar con experiencia previa en el equipo, WordPress se erigía como la opción más clara.

Gracias a su generosa cuota de mercado, se observaron varias vías para obtener las especificaciones necesarias para adaptar el sitio web a los requerimientos, siendo dos de ellos clave: la posibilidad de servir la web en varios idiomas (español y portugués al menos) y mostrar datos en forma de tablas o de descargables (.csv, .xls).

Tras un estudio de plugins disponibles, se optó por WPML, un clásico del ecosistema WordPress, para los idiomas y el potente wpDataTables para la visualización de datos, ambos plugins resolviendo las problemáticas sin mayor dificultad.

Con dichas decisiones tomadas, el siguiente paso era estudiar las mejores opciones para unificar y servir los datos a dicho gestor.

Originariamente se estudió trabajar con datos propios, tanto de las Islas Canarias como de Madeira y de Azores; pero dadas las asimetrías de actualizaciones y formatos se intentó desde un primer momento buscar una alternativa.

Por suerte, ya existe un sitio web con datos de las regiones NUTS de la Macaronesia, actualizados y de libre acceso, éste es el Portal Web del Eurostat.

Tras hacer la selección mano a mano de todas las tablas que podrían ser de interés estadístico, y que también tuvieran datos disponibles para los para los archipiélagos, comenzaron las pruebas. Se detectaron los siguientes problemas:

1. Aunque las operaciones usaban activos semánticos para codificar las diferentes variables de los datasets, estos activos no existían en Infraestructura de Datos y Metadatos Estadísticos del ISTAC (eDatos).
2. Estos activos semánticos sólo estaban disponibles en inglés.
3. El plugin para la visualización de datos sólo admite ficheros de tipo JSON cuando se trata de fuentes externas. Sin embargo, el Eurostat ofrece los datos en archivos con formato CSV.

Para el primer y el segundo problema se planteó una solución que aunque inicialmente pudiese parecer tediosa, resultaba bastante práctica y además permitiría crecer con facilidad. Esta solución consistía en crear una suerte de "decodificador" que contenga las clasificaciones en inglés y sus traducciones en español y portugués.

Para ello, fue necesario revisar a mano las 58 tablas, conteniendo 1257 códigos recogidos en 46 clasificaciones diferentes. Con esta información se pudo elaborar una tabla maestra en la que se recogían todos los códigos con sus correspondientes literales en español y en portugués.

El resultado fue una tabla similar a la siguiente:

| cl | code | ES | PT |
|---|---|---|---|
| accommod | BEDRM | Habitaciones | Quartos |
| accommod | BEDPL | Plazas de cama | Lugares de cama |
| +1255 filas | ... | ... | ... |

Esa tabla nos permitía tener un códice listo para integrar en la automatización que resolvía también el tercer problema, el de utilizar datos JSON en lugar de CSV.

La metodología de automatización se puede resumir en cuatro grandes etapas: scraping de datos desde la web de Eurostat, preparación de los datos, almacenamiento y notificación.

La web de Eurostat publica información estadística sobre territorios de la Macaronesia pero los datos están distribuidos por distintas páginas y secciones. Es por ello que se hizo necesario disponer de un repositorio de urls desde las que descargar los datasets. A este efecto se preparó una hoja de cálculo en Google Drive con metadatos asociados a cada fichero y una columna con la url al dataset. Haciendo uso de un paquete propio de Python desarrollado al efecto yagdrive se consulta dicho documento y se descarga la información

como un dataset de Pandas. Este dataset contiene, entre otras, información sobre codelists, traducciones o disponibilidad de los datos. Para cada una de las filas de este dataset se descarga el conjunto de datos asociado como ficheros .csv y se realiza una preparación del mismo.

La preparación del fichero de cada conjunto de datos SIEMAC consta de los siguientes cinco pasos:

1. Carga del .csv como un dataframe de Pandas
2. Filtrado de los registros
3. Limpieza de los datos
4. Recodificación
5. Exportación.

En primer lugar se filtran los registros atendiendo a los códigos geográficos de las regiones de la Macaronesia, a continuación se limpia el dataset de valores nulos y se reemplazan las comas por puntos en los valores de tipo numérico.

El siguiente paso es la recodificación de los datos para los idiomas indicados (en este caso español y portugués) utilizando la tabla de traducciones indicada anteriormente.

Por último se almacena el dataframe de salida tanto en formato .tsv como en formato .json. Finalmente quedará en local una estructura de ficheros con la siguiente nomenclatura:

```
├agr_r_accts_es.json
├agr_r_accts_es.ts
├agr_r_accts_pt.json
└agr_r_accts_pt.tsv
```

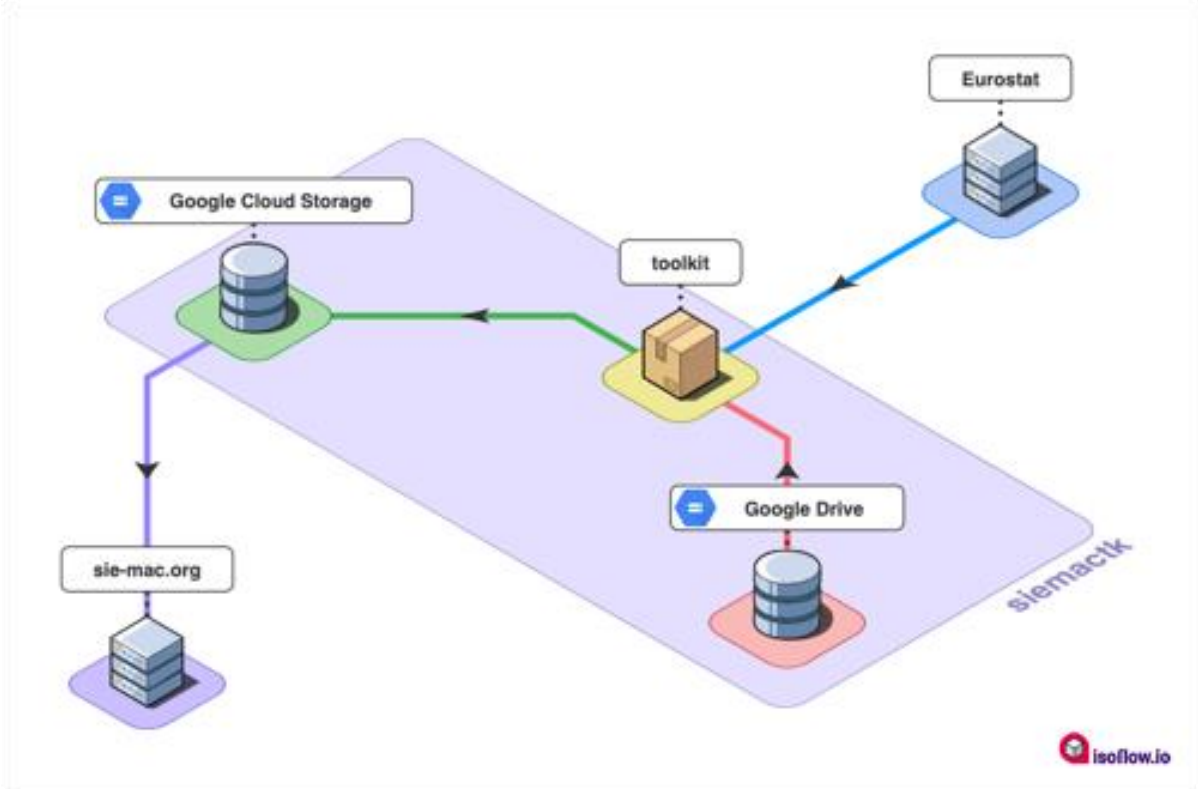
agr_r_accts es el identificador del conjunto de datos asociado a Cuentas económicas de la agricultura, y es sólo un ejemplo que se repite para el resto de indicadores asociados.

Para que todos estos ficheros estén disponibles es necesario alojarlos en algún servicio de la nube que sea fiable y de acceso rápido. Estamos hablando de más de 200 ficheros. Para ello se decidió utilizar Google Cloud Storage (GCS). En esta etapa se sube cada fichero a un bucket de GCS y se obtiene una url de descarga.

La última etapa en la automatización consiste en notificar por correo electrónico el listado de ficheros disponibles enlazados a cada una de sus ubicaciones GCS, de tal manera que se simplifica enormemente el acceso a los mismos.

Todo el proceso de automatización está diseñado como una tarea programada que se ejecuta el día 1 de cada mes a las 7 de la mañana a través de GitHub Actions, de forma totalmente autónoma y sin ninguna intervención humana.

En la siguiente imagen puede verse un resumen del funcionamiento general del proceso de automatización para la carga de datos:




## Resultados

Finalmente el producto entregado es una web autosostenible en materia de datos, con puntuales actualizaciones en la parte frontal (plugins, actualizaciones del núcleo de WordPress, el propio plugin del visualizador, etc.)

En una primera versión de la herramienta se utilizaba un fichero .csv en local para cargar las urls de los datasets de Eurostat. Esta aproximación tenía el inconveniente de que si los técnicos estadísticos querían añadir nuevas fuentes había que actualizar el fichero local. Es por ello que se decidió leer directamente de la hoja de cálculo Google Drive para obtener la información de partida. Esto mejoró notablemente el uso de la herramienta ya que se desacopla la entrada y permite que la hoja de cálculo se actualice de manera independiente a las ejecuciones.

- En la última ejecución: 232 ficheros generados (116 .json y 116 .tsv)
- Duración media las ejecuciones: 7 minutos
- 75% de ejecuciones correctas.

## Conclusión

Hacer uso de los datos ya publicados por Eurostat, y que a su vez son transmitidos por las oficinas estadísticas a dicho organismo, ha permitido partir de una fuente común con información en el mismo formato. De esta forma, se evita el tener que integrar datos desde diferentes sistemas de información. Esta ventaja no sólo aceleró el desarrollo actual sino que también permitirá que en un futuro próximo se añadan datos de Cabo Verde sin apenas esfuerzo.

Una vez seleccionado el origen de datos a utilizar, el siguiente reto estaba en conseguir que los datos se publicasen en los idiomas oficiales de las regiones de la Macaronesia (Español y Portugués) y en simplificar el proceso de carga para que a pesar de hacerse con una elevada frecuencia, no conllevase una carga de trabajo adicional para el personal del ISTAC.

En este sentido, la automatización completa del proceso de obtención, normalización, traducción y difusión de los datos ha permitido cumplir y superar las expectativas que inicialmente se tenían.

Por último, comentar que el uso de herramientas como Google Drive combinado con la implementación llevada a cabo para la automatización ha supuesto una gran ventaja porque se consigue dispone de un procedimiento genérico en el que la implementación no esté atada a las fuentes que se quieren consultar.

## Principales referencias bibliográficas

[1] Portal sie-mac.org - https://sie-mac.org/

[2] Descripción del proyecto SIE-MAC en el marco del programa INTERREG III-B - https://sie-mac.org/proyectos/sie-mac/

[3] YagDrive Documentation - https://yagdrive.readthedocs.io/en/latest/

[4] Repositorio siemactk - https://github.com/eDatos/siemactk

[5] Google Cloud Storage - https://cloud.google.com/storage

[6] Eurostat Database - https://ec.europa.eu/eurostat/data/database