Islas Canarias
Del 15 al 19 de noviembre de 2021

# Librerías R y Python para acceder a la API del ISTAC

**Sergio Delgado Quintero**
Instituto Canario de Estadística (ISTAC)
sdelquin@gmail.com

**Luis Pérez Geraldino**
Ecosistemas Virtuales y Modulares (EVM)
luis.perez@evm.net

**PALABRAS CLAVE**

API, conectores, datos abiertos, R, Python

# INTRODUCCIÓN

El movimiento Open Data es un esfuerzo para hacer que cada vez más información esté accesible y disponible públicamente para lograr mayor transparencia sin restricciones técnicas o legales, incluyendo también a administraciones y organismos públicos. No obstante, en muchos casos los datos proporcionados no están en un formato estandarizado y manejable, por lo que surge la necesidad de adaptar el código empleado para el tratamiento de los datos a fuentes de datos específicas para contemplar las posibles variaciones en los formatos de los datos en crudo, acceder a detalles para que los usuarios finales puedan evitar tareas de programación repetitivas y ahorrar tiempo mediante la simplificación, estandarización y automatización de los flujos de trabajo en análisis que faciliten la reproducción, el uso compartido de código y el análisis de datos.

En este contexto, aun siendo la interfaz gráfica de usuario (GUI) basada en web la herramienta principal usada por los usuarios para la consulta y descarga de los datos en diversos formatos, el Instituto Canario de Estadística (ISTAC) dispone de un catálogo de APIs abiertas[1] en el que se pone a disposición de cualquier usuario una gran cantidad de recursos en formato abierto, producto de la producción estadística del organismo. El catálogo, en el momento de la escritura de este documento, está estructurado en 9 grandes bloques:

1. **Indicadores.** Medidas de la intensidad de un fenómeno en el espacio-tiempo, con distintas granularidades espaciales y/o unidades temporales. Se estructuran en sistemas de indicadores, diseñados para aportar información más precisa sobre un fenómeno, organizándose en dimensiones.
2. **Recursos estadísticos.** Productos que se obtienen de una operación estadística.
3. **Recursos estructurales.** Recursos que sirven de apoyo para normalizar los recursos estadísticos a nivel estructural.
4. **Operaciones estadísticas.** Inventario de operaciones estadísticas del Instituto, así como las agrupaciones en familias e instancias de las operaciones.
5. **Registro SDMX.** Artefactos del estándar SDMX que el ISTAC utiliza y tiene disponibles en sus sistemas, para centralizar la localización de metadatos estadísticos y se pueda reutilizar por terceros.
6. **Metadatos comunes.** Inventario de la configuración de metadatos comunes e identificados como reutilizables dentro de la organización.
7. **Exportaciones.** Funcionalidad necesaria para poder exportar los recursos estadísticos en diferentes formatos.
8. **Callejero (Gazetteer).** Servicio de diccionario geográfico basándose en los diferentes municipios de las Islas Canarias.
9. **Enlaces permanentes.** Permalinks generados para que una URL sea accesible a lo largo del tiempo.

Cada uno de los bloques del catálogo cuenta con una serie de secciones. Para cada sección, a su vez, se definen una serie de puntos de entrada, que corresponde con cada una de las URLs disponibles.

[1] https://datos.canarias.es/api/estadisticas/

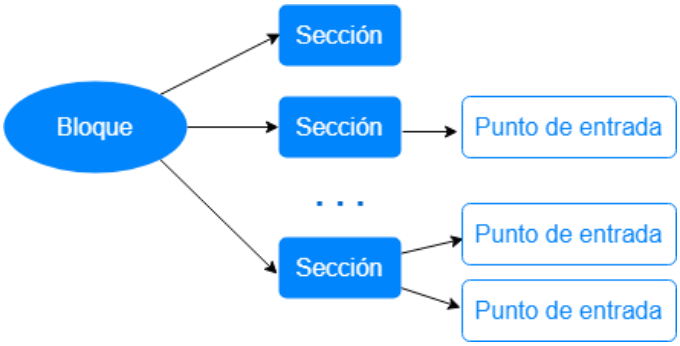


**Figura 1. Jerarquía de elementos del catálogo de datos abiertos**

Debido a la arquitectura y el diseño del catálogo de APIs del ISTAC, la mayoría de las peticiones a los recursos disponibles requieren de un conjunto de argumentos que, o bien son decisión del usuario, o bien provienen de otras llamadas previas.

A su vez, la respuesta que ofrece la API para cada petición se compone, en la mayor parte de las ocasiones y como opción por defecto, en formato JSON (JavaScript Object Notation). El formato JSON es un formato abierto utilizado como alternativa al XML para la transferencia de datos estructurados, cuyo uso se ha extendido en poco tiempo gracias a su sencillez y facilidad de uso e implementación para todos los lenguajes de programación. Un JSON no es más que una cadena cuyo formato recuerda al de los objetos literales JavaScript, permitiendo incluir diferentes tipos de datos como cadenas, número, arreglos, booleanos, y otros literales de objeto, lo cual permite construir una jerarquía de datos en distintos niveles.

A pesar de ser un formato ligero de intercambio de datos, que resulta sencillo de leer y escribir para los programadores y simple de interpretar y generar para las máquinas, siendo independiente del lenguaje de programación que se emplee para trabajar con él, resulta un tanto engorroso cuando lo que se desea es obtener valor de la información que contiene.

Dicho JSON ofrece una gran cantidad de metadatos, además de los propios datos consultados. Simplificando, podríamos decir que existe un listado de dimensiones y un conjunto de datos asociados a esas dimensiones.

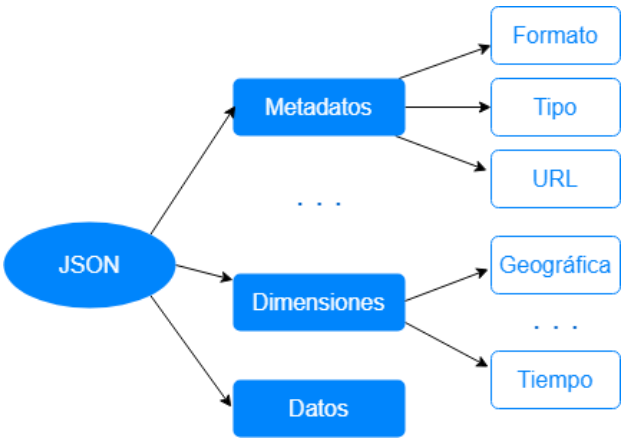


**Figura 2. Estructura de una respuesta de la API**

Dado que la eficiencia en los sistemas de información está basada en la velocidad de procesamiento, la confiabilidad, la seguridad y el aprovechamiento de los recursos, la disposición de los datos y el manejo adecuado del espacio de memoria es necesario para el mejoramiento de los procesos, por lo que es muy importante diseñar y utilizar estructuras de datos coherentes para facilitar su uso. Con ello en mente y el objetivo de facilitar la utilización de los datos, y que la respuesta no ocupe un tamaño excesivamente grande (con la mejora en tiempos de red que ello conlleva), para el catálogo de datos se ha diseñado una respuesta donde los datos vienen listados en un orden que corresponde al producto cartesiano de las granularidades de las dimensiones.

Se entiende que la calidad de un análisis de datos depende en gran medida de los datos en los que se basa. A menudo las distintas herramientas de análisis ofrecen conectores para facilitar la lectura e interpretación de los datos, independientemente de la estructura y el formato en el que se encuentren. Es por ello que desde el ISTAC se proporcionan una serie de conectores a modo de intermediación para la consulta de datos propios, como por ejemplo PowerBI, Google Data Studio, QGIS o Tableau. La finalidad de estos conectores no es otra que la de comunicarse internamente con la API, realizando consultas y posteriormente dar un formato más amigable a las respuestas obtenidas. Ofrecer este tipo de herramientas ofrece una ventaja para el usuario final, ya que se accede de manera sencilla, organizada y estructurada a los datos disponibles en el catálogo publicado.

Cada conector tiene sus peculiaridades de desarrollo, a nivel tecnológico o de requisitos de las distintas herramientas de análisis de datos, pero la filosofía a seguir se mantiene durante el conjunto de conectores. Actualmente no existen mecanismos para facilitar la lectura de los datos disponibles en el catálogo del ISTAC para los lenguajes R y Python, dos lenguajes de programación que se han convertido en estos últimos años, por popularidad, en estándares para el ámbito de la ciencia de datos.

## OBJETIVOS

Con la creación de las librerías R y Python se persigue, como principal objetivo, simplificar el acceso a los datos proporcionados por el catálogo de datos abiertos del ISTAC. De esta forma, se consigue eliminar la posible complejidad inicial de compresión de la estructura de las respuestas obtenidas del catálogo de datos, ya que con la forma que tiene de encapsular los datos no es sencilla a priori y, aunque tiene sus ventajas, no llega a ser del todo intuitiva para el usuario que no está familiarizado con estructuras similares. Además, gracias a la creación de las mismas, se facilita la obtención de objetos en formatos estructurados típicos utilizables inmediatamente por librerías en los lenguajes de programación ampliamente empleados dentro del mundo de la ciencia de datos, como lo son R y Python, además de fomentar la reutilización de la información de forma coherente e intuitiva.

Así mismo, se considera oportuno proporcionar una interfaz más simplificada para el acceso a determinadas secciones del catálogo de datos abiertos. Esta interfaz, que si bien no es completa en términos funcionales, sí facilita el manejo de las librerías por usuarios no iniciados. De forma que aquellos que tengan interés en realizar análisis de ciertos datos en específico, sean capaces de encontrarlos de forma rápida y sencilla, en un entorno que les permita no solo obtener el dato sino también realizar el análisis haciendo uso de una misma herramienta para ambas cosas.

Otro de los objetivos a medio plazo es eliminar la necesidad de consulta directa a la API cuando se quiere obtener algún dato que se encuentre en el catálogo. No se trata de imponer el uso exclusivo de la librería, simplemente ofrecer una alternativa con formatos más amigables y sencillos para algunos casos.

Finalmente, en consonancia con el conjunto de conectores corporativos ya existentes, se busca completar el conjunto de herramientas de lectura de catálogo de datos abiertos del ISTAC en los dos lenguajes de programación más populares en el ámbito de la ciencia de datos.

## METODOLOGÍA

En primer lugar las librerías ofrecen una serie de funciones, debidamente documentadas, que se vinculan directamente a los puntos de entrada de la API. Estas funciones admiten parámetros similares a lo que sería una petición directa a la API. Una respuesta típica de la API puede tener este aspecto:


```
{
    "kind": "indicators#indicatorData",
    "selfLink": "https://datos.canarias.es/api/estadisticas/indicators/v1.0/indicators/EMPLEO_REGISTRADO_HOSTELERIA/data",
    [...],
    "format": [
        "GEOGRAPHICAL",
        "TIME",
        "MEASURE"
    ],
    "dimension": {
        "GEOGRAPHICAL": {
            "representation": {
                "size": 96,
                "index": {
                    "ES70": 0,
                    "35004": 1,
                    [...]
                }
            }
        },
        "TIME": {
            "representation": {
                "size": 62,
                "index": {
                    "2021-Q2": 0,
                    "2021-Q1": 1
                    [...]
                }
            }
        },
        "MEASURE": {
            "representation": {
                "size": 1,
                "index": {
                    "ANNUAL_PERCENTAGE_RATE": 0
                }
            }
        }
    },
    "observation": [
        "-7.12",
        "1.43",
        "-9.29",
        "-3.91",
        [...]
    ]
}
```


METADATOS

DIMENSIONES

DATOS

Se plantea realizar las transformaciones necesarias para que los datos tengan un formato más amigable y manejable para el usuario, como listados o dataframes tanto en Python como en R. Lo que se plantea es obtener algo similar a:

| DIMENSIONES | | | DATOS |
|---|---|---|---|
| **geographical** | **time** | **measure** | **observation** |
| ES70 | 2021-Q2 | ANNUAL_PERCENTAGE_RATE | -7.12 |
| ES70 | 2021-Q1 | ANNUAL_PERCENTAGE_RATE | 1.43 |
| ... | ... | ... | ... |
| 35004 | 2021-Q2 | ANNUAL_PERCENTAGE_RATE | ... |
| 35004 | 2021-Q1 | ANNUAL_PERCENTAGE_RATE | ... |
| ... | ... | ... | ... |

Se ha realizado una planificación para implementar por completo las funciones necesarias para consultar la totalidad de datos publicados en el catálogo de datos abiertos, aunque se irá implementando por módulos para poder priorizar el acceso a aquellos conjuntos de datos que se consideren más relevantes.

## RESULTADOS

Como primer resultado tenemos la publicación de versiones de ambas librerías en los gestores de paquetes oficiales de R (https://cran.r-project.org/web/packages/istacr/index.html) y Python (https://pypi.org/project/istacpy/).

En esta versión se facilita el acceso al catálogo de datos de tres de los bloques que se publican:

- Indicadores
- Recursos estructurales
- Recursos estadísticos

A nivel general, se presta especial atención a la conversión de los datos obtenidos en las distintas llamadas en dataframes, con estructura similar pero nativa de cada lenguaje.

Específicamente, dentro del bloque de indicadores en la librería, cabe destacar el submódulo lite. Dada la naturaleza de un indicador, y basándose en un enfoque de programación orientada a objetos, es posible simplificar las operaciones de filtrado y selección de datos.

Se ha hecho un esfuerzo por proporcionar una interfaz flexible y simplificada utilizando únicamente la dimensión temporal y la dimensión geográfica. Fijando las granularidades de cada una de estas dimensiones obtenemos un dataframe con datos del indicador en cuestión. Así mismo se han añadido filtros de búsqueda como parámetros de las funciones que son aplicables a distintos niveles de granularidad.

Igualmente se ha generado documentación vinculada a cada proyecto:

- **Python:** https://istacpy.readthedocs.io/
- **R:** https://cran.r-project.org/web/packages/istacr/istacr.pdf

De la misma manera, se continuará con la planificación iniciada para acercar los conjuntos de datos publicados en el catálogo del ISTAC a los usuarios finales con la metodología que se propone.

## CONCLUSIÓN

En los últimos años han surgido una gran cantidad de herramientas que permiten compartir información en diversos formatos de open source a través de diferentes plataformas en la nube haciendo uso de servicios web como las APIs. Sin embargo, uno de los problemas derivados de ello es el cómo recopilar datos de interés de dichas fuentes.

Las librerías istacr e istacpy permiten a los usuarios consultar y obtener series y conjuntos de datos estadísticos de manera rápida y eficiente. El mayor atractivo de estas librerías consiste en conectarse al catálogo abierto de APIs del ISTAC para acceder a los datos y, luego, crear un conjunto de datos en R o Python para trabajar con él y extraer valor de los mismos. Lo cual representa una ventaja frente a otros procedimientos que implican la descarga de ficheros en local para su posterior carga, tratamiento y análisis.

## BIBLIOGRAFÍA


1. VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.".
2. McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".
3. Kabacoff, R. I. (2015). R in action: data analysis and graphics with R. Simon and Schuster.