DESCRIPCIÓN FLUJO DE LA FUNCIÓN AZUL

En el presente documento se describe los procesos realizados en la Función Azul.

En este flujo de procesamiento se utilizan como inputs los siguientes parámetros:

* **Ruta\_Nombre\_Input\_entidades**  ("DATA\_SIC\_ENTIDADES\_PUBLICAS\_SOLICITUD\_ESP.csv")
* **Ruta\_Nombre\_Input\_investigadores** ("DATA\_SIC\_INVEST\_UO\_FIL\_PB\_J\_MAESTRO.csv")
* **Ruta\_Nombre\_Input\_relaciones** ( "DATA\_SIC\_RELACIONES\_ENTIDADES\_PB\_AYUDAS\_ESP\_A6CORREC\_new.csv")
* **Ruta\_Nombre\_Input\_invente** ('Analisis\_entidades\_invente\_15-02-2022.xlsx')
* **Ruta\_Nombre\_Input\_dir3** ("DIR3\_concatenado\_AGE\_CCAA\_EELL\_Universidades.csv")
* **Ruta\_Nombre\_Intermedio\_cruce1\_relaciones** ("DATA\_SIC\_NV\_RELACIONES\_MST"
* **Ruta\_Nombre\_Intermedio\_cruce1\_entidades** ("DATA\_SIC\_ENTIDADES\_CENTROS\_PB"
* **Ruta\_Nombre\_Intermedio\_cruce2\_invente** ("test/DATA\_SIC\_ENTIDADES\_CENTROS\_PB\_INVENTE\_NORM")
* **Ruta\_Nombre\_Intermedio\_cruce3\_dir3** ("test/DATA\_SIC\_ENTIDADES\_CENTROS\_PB\_INVENTE\_DIR3\_NORM")
* **Ruta\_Nombre\_Output\_investigadores** ("DATA\_SIC\_INVEST\_UO\_FIL\_PB\_CENTRO\_MST")
* **Ruta\_Nombre\_Output\_relaciones** ("DATA\_SIC\_RELACIONES\_ENTIDADES\_PB\_FIL\_INVEST")
* **Ruta\_Nombre\_Output\_entidades** ("DATA\_SIC\_ENTIDADES\_PUBLICAS\_FIL\_INVEST")
* **flag\_invente\_dir3** (True o False)

1. Si el valor es True, se crean diferentes cruces entre los archivos, completando el flujo 1.
2. Si es False se procesa el flujo 2.

* **Flag\_csv** (True o False)

1. Cuando sea True se generan los archivos en los formatos csv y parquet.
2. Cuando sea False, se generan los archivos únicamente en formato parquet.

Esta función consta de siete (07) subprocesos descritos a continuación:

1. **Función cruce1\_entidades\_Output\_investigadores:** Esta función almacena en la rutas establecidas los archivos correspondientes a Nombre\_Intermedio\_cruce1\_entidades y Nombre\_Output\_investigadores.

Y su proceso es el siguiente:

**Archivo Nombre\_Intermedio\_cruce1\_entidades**: debe tener las mismas columnas que el archivo **Nombre\_Input\_entidades**. Vamos a incluir en este archivo la información de los centros, de aquellos registros cuyo binomio entidad-centro no está incluido (no han superado el valor del threshold) en el archivo **Nombre\_Input\_relaciones**, como entidades en este archivo. Para ello, se rellenan las columnas de la siguiente manera:

* Incluimos en la columna Id del archivo **Nombre\_Intermedio\_cruce1\_entidades** números enteros para los nuevos registros obtenidos. Se debe continuar la numeración a partir del último valor que tiene esta columna en el fichero Nombre\_Input\_entidades.
* Incluimos en la columna Entidad\_Norm los nombres de los centros (columna Centro\_Norm del archivo Nombre\_Input\_investigadores).
* Incluimos en la columna Provincia\_Entidad, las provincias de los centros (columna Provincia\_Centro del archivo Nombre\_Input\_investigadores).
* Añadimos a las columnas List\_Entidad\_Norm y List\_Provincia\_Entidad los datos de las columnas Centro\_Norm y Provincia\_Centro, que tenemos en el archivo Nombre\_Input\_investigadores.
* List\_CIF se mantiene vacío para estos registros al no tener CIF asociado a los centros.

IMPORTANTE: eliminar registros duplicados usando las columnas **Entidad\_Norm** y **Provincia\_Entidad**, ya que podemos tener nombres de centros similares. Tener en cuenta la similitud del nombre en la columna **Entidad\_Norm**.

El archivo **Nombre\_Intermedio\_cruce1\_entidades** se guarda en la ruta, con el nombre especificado en el input, en formato csv y parquet (dependiendo del valor de flag\_csv).

**Archivo Nombre\_Output\_investigadores:** tiene las mismas columnas y registros que el archivo Nombre\_Input\_investigadores. Hacemos aquí una diferenciación:

Para aquellos registros del archivo **Nombre\_Input\_investigadores**, cuyo binomio entidad-centro ya está contenido en el archivo **Nombre\_Input\_relaciones** (han superado el threshold que es 0.875), rellenamos la información de las siguientes columnas del archivo **Nombre\_Output\_investigadores**:

* Se incluye toda la información que tenemos en **Nombre\_Input\_investigadores** para estos registros
* En la columna del **Id\_Centro** incluimos los identificadores **Id\_Nivel\_1** del archivo **Nombre\_Input\_relaciones**,
* En la columna **Centro\_Nombre\_match** incluimos el nombre del centro (**NOMBRE\_ENTIDAD\_NIVEL\_1** del archivo **Nombre\_Input\_relaciones** que ha sido previamente normalizada)
* En la columna **Centro\_Provincia\_match** incluimos la provincia (columna **COD\_PROVINCIA\_NIVEL\_1**) que se encuentran en el archivo **Nombre\_Input\_relaciones**.
* En la columna **Centro\_match**, que es donde se indica como han cruzado, especificamos “**Nombre Centro Invest Relaciones**”.
* Creamos la columna **Similitud\_Centro** que contiene los valores output de la función Ratcliff-Obershelp (Final\_Score), que ya habíamos calculado usando las columnas **Centro\_Norm** del archivo **Nombre\_Input\_filiaciones** y **NOMBRE\_ENTIDAD\_NIVEL\_1** normalizada del archivo **Nombre\_Input\_relaciones**.
* La posición que deben tener las nuevas columnas creadas se indica en la siguiente figura:
* 
* Para aquellos registros del archivo **Nombre\_Input\_investigadores** cuyo centro se haya incluido como nueva entidad en el archivo **Nombre\_Intermedio\_cruce1\_entidades** (registros que no han superado el threshold), rellenamos la información de las siguientes columnas del archivo **Nombre\_Output\_investigadores**:
* En la columna **Id\_Centro** incluimos el identificador Id asignado a ese centro en el archivo **Nombre\_Intermedio\_cruce1\_entidades**.
* Incluimos en la columna Similitud\_Centro (**Final\_Score** que será menor que el valor threshold) los valores output de la función Ratcliff-Obershelp, que ya habíamos calculado usando las columnas **Centro\_Norm** del archivo **Nombre\_Input\_filiaciones** y **NOMBRE\_ENTIDAD\_NIVEL\_1** normalizada del archivo **Nombre\_Input\_relaciones**.

El archivo **Nombre\_Output\_investigadores** se guarda en la ruta, con el nombre especificado en el input, en formato csv y parquet (dependiendo del valor de flag\_csv).

NOTA: Se crean las columnas **Id\_Centro**, **Centro\_Nombre\_Match**, **Centro\_Provincia\_Match**, **Centro\_Match** y **Similitud\_Centro**, sí y solo sí, no existen.

1. **Función cruce1\_relaciones:** Esta función alamacena y procesa los archivos correspondientes a **Nombre\_Intermedio\_cruce1\_relaciones**

Archivo Nombre\_Intermedio\_cruce1\_relaciones: debe tener las mismas columnas que el archivo Nombre\_Input\_relaciones. Estas columnas se rellenan como se indica a continuación, para los registros cuyos binomios entidad-centro no están en el archivo Nombre\_Input\_relaciones:

* En la columna **Id\_NIVEL\_0** incluimos el valor de la columna Id del archivo **Nombre\_Input\_investigadores**
* En la columna **Id\_NIVEL\_1** incluimos el valor de la columna **Id** que se corresponde con los nuevos centros que estamos incluyendo en el archivo **Nombre\_Intermedio\_cruce1\_entidades**.
* En la columna **NOMBRE\_ENTIDAD\_NIVEL\_0** incluimos el nombre de la entidad del archivo **Nombre\_Input\_investigadores** (columna Entidad),
* En la columna **NOMBRE\_ENTIDAD\_NIVEL\_1** incluimos el nombre del centro del archivo **Nombre\_Input\_investigadores** (columna Centro),
* En la columna **COD\_PROVINCIA\_NIVEL\_0** incluimos los datos de la columna Provincia\_Entidad del **Nombre\_Input\_investigadores**
* En la columna **COD\_PROVINCIA\_NIVEL\_1** incluimos los datos de la columna Provincia\_Centro del **Nombre\_Input\_investigadores**
* En la columna **NIF\_COD\_NIVEL\_0** incluimos los datos de la columna **CIF** del archivo **Nombre\_Input\_investigadores**.
* En la columna Jerarquía especificamos “**Orgánica Directa**”.

IMPORTANTE: Aquellos registros que tengan el mismo nombre (o uno similar con un valor de la similitud superior al threshold), en las columnas **Entidad\_Norm** y **Centro\_Norm** en el archivo **Nombre\_Input\_investigadores**, no serán incorporados en el archivo **Nombre\_Intermedio\_cruce1\_relaciones**. Ejemplo: Entidad\_Norm = ‘instituto de investigacion’, Centro\_Norm = ‘instituto de investigacion’.

El archivo **Nombre\_Intermedio\_cruce1\_relaciones** se guarda en la ruta, con el nombre especificado en el input, en formato csv y parquet (dependiendo del valor de **flag\_csv**).

1. **Función output\_relaciones:** Función utilizada para procesar y almacenar los archivos correspondientes a **Nombre\_Output\_relaciones**.

Los registros del archivo **Nombre\_Intermedio\_cruce1\_relaciones** son añadidos verticalmente, como nuevos registros, al archivo **Nombre\_Input\_relaciones**. A este archivo resultante lo denominamos **Nombre\_Output\_relaciones**.

El archivo **Nombre\_Output\_relaciones** se guarda en la ruta, con el nombre especificado en el input, en formato csv y parquet (dependiendo del valor de flag\_csv).

1. **Función cruce2\_invente:** Esta función procesa y almacena los archivos correspondientes a **Nombre\_Intermedio\_cruce2\_invente** cuando si flag\_invente\_dir3 == True.

Creamos el archivo **Nombre\_Intermedio\_cruce2\_invente**. Este archivo tiene inicialmente las mismas columnas y los mismos registros que el archivo **Nombre\_Intermedio\_cruce1\_entidades**. Para ello, estudiamos la similitud de nombre de los registros del archivo **Nombre\_Intermedio\_cruce1\_entidades** con **Nombre\_Input\_invente** solo para complementar la información (si entran 100 registros, tras el cruce con Invente salen 100 registros). Este estudio se realiza utilizando las columnas **Entidad\_Norm** del archivo **Nombre\_Intermedio\_cruce1\_entidades** y **DenominacionSocial** de Nombre\_Input\_invente (previamente normalizada). Para aquellos registros que tengan una similitud superior al valor threshold, se completa la información de las columnas del archivo Nombre\_Intermedio\_cruce1\_entidades referentes a **Nombre\_Input\_invente** que estaban vacías antes del cruce. Estas columnas son: **CodigoInvente**, **DenominacionSocial**, **FormaJuridica\_Codigo**, **FormaJuridica\_Descripcion**, **NIF**, **codigoDir3**, **codigoOrigen** y **Provincia\_Codigo**. Además, incluimos la siguiente información en las columnas del archivo **Nombre\_Intermedio\_cruce2\_invente** que se especifican:

• En la columna **Entidad\_Norm** incluimos la información de la columna **DenominacionSocial** (previamente normalizada),

• En la columna **Provincia\_Entidad** incluimos la información de la columna **Provincia\_Codigo**,

• En la columna **CIF** incluimos la información de la columna **NIF**.

• Creamos la columna **Similitud\_invente**. En ella incluimos los valores de similitud (**Final\_Score**) obtenidos al aplicar la función de similitud entre las columna **Entidad\_Norm** del archivo **Nombre\_Intermedio\_cruce1\_entidades** y **DenominacionSocial** del archivo **Nombre\_Input\_invente.**

El archivo resultante lo denominamos **Nombre\_Intermedio\_cruce2\_invente** y se guarda en la ruta y con el nombre especificado en el input, en formato csv y parquet (dependiendo del valor de **flag\_csv**).

1. **Función cruce3\_dir3:** Esta función procesa y almacena el archivo correspondiente a Nombre\_Intermedio\_cruce3\_dir3.

Creamos el archivo **Nombre\_Intermedio\_proceso\_dir3\_entidades**. Este archivo tiene inicialmente las mismas columnas y los mismos registros que el archivo **Nombre\_Intermedio\_cruce2\_invente**. Para ello, estudiamos la similitud de nombre de los registros del archivo **Nombre\_Intermedio\_cruce2\_invente con Nombre\_Input\_dir3** solo para complementar la información (si entran 100 registros, tras el cruce con **Nombre\_Input\_dir3** salen 100 registros). Este estudio se realiza usando las columnas **Entidad\_Norm** del archivo **Nombre\_Intermedio\_cruce2\_invente** y **C\_DNM\_UD\_ORGANICA** (la cual ha sido previamente normalizada) del archivo **Nombre\_Input\_dir3**. Debemos tener en cuenta que:

• Si una misma entidad aparece en varios registros del cruce (es decir, tenemos varios registros que tienen un valor de similitud por encima del threshold entre las columnas Entidad\_Norm del archivo Nombre\_Intermedio\_cruce2\_invente y la columna **C\_DNM\_UD\_ORGANICA** del archivo **Nombre\_Input\_dir3**), y estos tienen distintos valores en la columna **N\_NIVEL\_JERARQUICO**, cogemos la información del registro que tenga el valor más pequeño en esta columna.

• Si una misma entidad aparece en varios registros del cruce (es decir, tenemos varios registros que tienen un valor de similitud por encima del threshold entre las columnas **Entidad\_Norm** del archivo **Nombre\_Intermedio\_cruce2\_invente** y la columna **C\_DNM\_UD\_ORGANICA** del archivo **Nombre\_Input\_dir3**), y además comparten el valor más bajo de la columna **N\_NIVEL\_JERARQUICO**, para esos registros cruzamos la columna **Entidad\_Norm** del fichero **Nombre\_Input\_investigadores** de ese registro, con la columna **C\_DNM\_UD\_ORGANICA\_SUPERIOR** del fichero **Nombre\_Input\_dir3**. Aquel registro de **Nombre\_Input\_dir3** con el que finalmente cruce, será del que nos traigamos la información.

Teniendo en cuenta estos dos puntos, se completa la información de las columnas del archivo **Nombre\_Intermedio\_cruce2\_invente** referentes a **Nombre\_Input\_dir3** (que estaban vacías antes del cruce). Además, completamos la información de las siguientes columnas:

• En la columna **Entidad\_Norm** incluimos la información de la columna **C\_DNM\_UD\_ORGANICA** (previamente normalizada)

• En la columna **Provincia\_Entidad** incluimos la información de la columna **C\_DESC\_PROV**

• En la provincia **CIF** incluimos la información de la columna **NIF\_CIF**.

• Creamos la columna **Similitud\_Dir3**. En ella incluimos los valores de similitud obtenidos entre las columna Entidad\_Norm del archivo **Nombre\_Intermedio\_cruce2\_invente** y **C\_DNM\_UD\_ORGANICA** del archivo **Nombre\_Input\_dir3**.

Al archivo resultante lo denominamos **Nombre\_Intermedio\_cruce3\_dir3** y se guarda en la ruta y con el nombre especificado en el input, en formato csv y parquet (dependiendo del valor de flag\_csv).

1. **Función Output\_entidades\_true:** Esta función procesa y almacena el archivo correspondiente a **Nombre\_Output\_entidades.**

Incluimos los registros del archivo **Nombre\_Intermedio\_cruce3\_dir3** en el maestro **Nombre\_Input\_entidades**. Para ello:

• eliminamos las columnas **Similitud\_Invente** y **Similitud\_Dir3** que tenemos en el archivo **Nombre\_Intermedio\_cruce3\_dir3**.

• concatenamos verticalmente los dos archivos.

El archivo resultante es denominado **Nombre\_Output\_entidades** y es una nueva versión del maestro de centros. Este archivo se guarda en la ruta y con el nombre especificado en el input, en formato csv y parquet.

1. **Función Output\_entidades\_false:** Esta función procesa y almacena el archivo correspondiente a **Nombre\_Output\_entidades** flag\_invente\_dir3== False.

Incluimos los registros del archivo **Nombre\_Intermedio\_cruce1\_entidades** en el archivo **Nombre\_Input\_entidades**. Hay que recordar aquí que el archivo **Nombre\_Intermedio\_cruce1\_entidades** ya tiene la misma estructura que el **Nombre\_Input\_entidades**, con lo cual directamente hacemos una concatenación vertical.

El archivo resultante es denominado **Nombre\_Output\_entidades** y se guarda en la ruta y con el nombre especificado en el input, en formato csv y parquet (dependiendo del valor de flag\_csv).