**Repositorio:** [**https://github.com/LMCappelletti/data.git**](https://github.com/LMCappelletti/data.git)

Trabajo sobre los tópicos <https://github.com/argendatafundar/data> descargados el 30/05/2025

**Post 29/05/2025 Reunión Proceso Renaming**

- en v1\_renaming.ipynb considerar el input country\_keys: nombres de columnas que refieren a algo geo referencial: <https://github.com/argendatafundar/geonomencladores-interno/blob/main/cosas/country_keys.txt> y con eso veo posibles columnas que v2\_renaming.ipynb no está tomando como georreferenciales pero que si lo son

- en v1\_renaming.ipynb hice una v2 de country\_keys.txt mergeando lo que la condición de % encuentra como georreferencial pero no está en country\_keys.txt. Esta nueva versión es new\_GeoRefsKeys.txt

- input, geonomenclador, es la pestaña front de (hacer copia para que la permita descargar):

[https://docs.google.com/spreadsheets/d/1fAQ3rmE3Fr6ztUKXvDYJw0iGH-Vidrx8wYkkHEgBBIQ/edit?gid=1603974855#gid=1603974855](https://docs.google.com/spreadsheets/d/1fAQ3rmE3Fr6ztUKXvDYJw0iGH-Vidrx8wYkkHEgBBIQ/edit?gid=1603974855" \l "gid=1603974855)

Uso la versión que modificó Nacho post prueba Tópico Pescas: llamada geonomenclador\_PostPesca.csv en el repositorio

**29/05/2025 Reunión Proceso Renaming: definiciones para seguir trabajando**

- al corto plazo es para lo que tenemos con lo ya hecho de tópicos argendata

- si el csv no tiene columna asociada a georreferencia no agregarles las columnas geocodigoFundar geonombreFundar a todos los csv nuevos

Agunas definiciones para los tipos de problemas:

df\_tipo1

composicion\_exportaciones\_bienes\_regiones\_mundo, para elegir pais usa el iso3

df\_tipo2

busca codigo y topico\_codigo para topicodeshum

+ACI-dataset tiene el codigo 'ZAR' que NO esta incluido en los geocodigos y tiene la string 'RD CONGO' que NO matchea sin ambiguedad en desc+AF8-fundar. Puede implicar modificaciones del geonomenclador (a revisi+//3//Q-n

me genero geocod acecon\_zar, hablo con Nacho

df\_tipo3

- varios geocodigos de desc\_fundar aparecen como topico\_geocodigo y en el cvs esta el gódico sin el topico\_ delante

- muchos owid que no estan en el geo

code: "OWID\_YAR", name: "Yemen Arab Republic"

hablarlo con nacho a ver que hacemos, agregamos codigo?

Lo de owid lo vi acá:

<https://github.com/owid/etl/blob/master/etl/steps/data/garden/regions/2023-01-01/regions.yml>

df\_tipo4

Tengo string sin codigo, lo hago a mano. Anotar tablita.

va a haber csvs que no cumplir

**Hasta 28/05/2025:**

**El proceso de renaming tiene como uno de sus productos un dataframe (df\_problemas) compuesto por las columnas: TOPICO | csv | problema**, pero puede haber porblemas que no sean detectados por el código por diversas razones.

**v2\_renaming.ipynb:**¿Cómo busco las columnas de interés para esta iteración?

- X% (por ejemplo 60%) del contenido de columnas de interés de un .csv de un tópico es parte de alguna de las columnas del geoneomenclador, en particular columnas del csv que se correspondan en un X% a “geocodigo” y “desc\_fundar” del geonomenclador

**El proceso de renaming considera lo siguiente** para rellenar las columnas 'geocodigoFundar' y 'geonombreFundar':

- **Si un dataset (cada csv original de cada topico) tiene una columna asociada a geocodigo** segun columnscsv\_Geocodigo\_descFundar, el contenido de la columna geocodigoFundar del csv copia tendrá el contenido de la columna asociada a geocodigo segun columnscsv\_Geocodigo\_descFundar. **La columna geonombreFundar del csv copia llenará cada una de sus filas acorde al name\_long (si es que existe), y si no es así al string de desc\_fundar correspondiente a ese geocodigo del geonomenclador.**

- **Si un dataset tiene una string que matchee sin ambigüedad en desc\_fundar del geonomenclador entonces se usa el código que está en el geonomenclador para esa string. Similar al caso anterior.**

- **En el caso de que un dataset tiene un código que NO esté incluido en los geocodigos pero tiene una string que matchee sin ambigüedad en desc\_fundar, entonces se usa el código que está en geocódigo en el geonomenclador.** Esto se registrará en df\_problemas con las columnas TOPICO | csv | problema: **"dataset tiene el codigo [...], que NO esta incluido en los geocodigos pero tiene la string [...] que matchea sin ambiguedad en desc\_fundar".** Por cada problema se guarda la string y el código correspondiente. Esto puede dar lugar que un mismo csv de un tópico tenga varios problemas a registrar. **NO HAY QUE MODIFICAR GEONOMENCLADOR, EL RENAMING ES HECHO A PARTIR DE LA COLUMNA desc\_fundar DEL GEONOMENCLADOR, pero quiero dejar registro.**

**- Si un dataset tiene un código que NO este incluido en los geocodigos y tiene una string que NO matchea sin ambigüedad en desc\_fundar, se genera un código nuevo.** Esto se registrará en df\_problemas con las columnas TOPICO | csv | problema: **"dataset tiene el codigo [...] que NO esta incluido en los geocodigos y tiene la string [...] que NO matchea sin ambiguedad en desc\_fundar. Puede implicar modificaciones del geonomenclador (a revisión, para no pisarse con otros casos)."**. Por cada problema se guarda la string y el código correspondiente. Esto puede dar lugar que un mismo csv de un tópico tenga varios problemas a registrar. **PODRÍA DAR LUGAR A MODIFICACIONES DEL GEONOMENCLADOR.**

**- Situacion similar a esta: Dataset tiene un código que NO está incluido en los geocódigos y no existe columna desc\_fundar**. En df\_problemas se reporta: **"dataset tiene el codigo [...] que NO esta incluido en los geocodigos y no existe columna desc\_fundar. Puede implicar modificaciones del geonomenclador (a revisión, para no pisarse con otros casos)."**

- **Si un dataset tiene una string que NO matchee sin ambigüedad en desc\_fundar**, entonces guardar este caso en el dataframe llamado df\_problemas con las columnas TOPICO | csv | problema, donde TOPICO es el topico con el que se corresponde, csv el csv evaluado y en problema poner: **"Para la string [...] se debería generar un codigo nuevo (a revisión, para no pisarse con otros casos)". PODRÍA DAR LUGAR A MODIFICACIONES DEL GEONOMENCLADOR.**

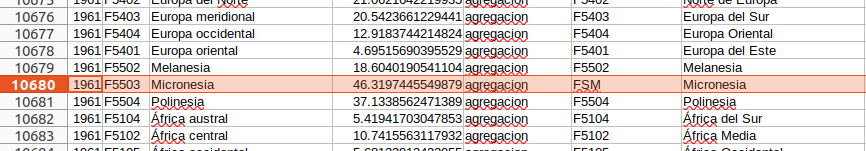
* Problemas que no aparecen en df\_problemas de renaming pesca:
  + Archivos donde no se hace el renaming mediante proceso automatizado:
    - **06\_desembarques\_puertos\_totales.csv**: porque se coloca Pcia. antes del nombre de cada provincia, el geonomenclador tiene el nombre directo provincia solución: sacar el Pcia.

**Ejemplos de tipos de problemas que si aparecen en df\_problemas de renaming pesca:**

*“dataset tiene el codigo 'F5503', que NO esta incluido en los geocodigos pero tiene la string 'MICRONESIA' que matchea sin ambiguedad en desc\_fundar”*

Entonces en el csv 31\_consumo\_per\_capita\_pescado\_anio\_pais.csv luego del renaming, la nueva columna geonombreFundar es *Micronesia* y su geocódigoFundar es *FSM* (como en el geonomenclador)

csv 31\_consumo\_per\_capita\_pescado\_anio\_pais.csv luego del renaming, la nueva columna geonombreFundar es *Micronesia* y su geocodigo





# PRUEBA SOBRE OTRO CSV DE UN TOPICO ARGENDATA CON PROBLEMA:

# TOPICO = 'MERTRA'

# ARCHIVO = 'lavarropas\_tasa\_empleo\_fem\_provincia.csv'

# Para la string 'CABA' generar un codigo nuevo (a revison para ser desambiguado con posibles matches)

# PRUEBA SOBRE OTRO CSV DE UN TOPICO ARGENDATA CON PROBLEMA:

# TOPICO = 'ACECON'

# ARCHIVO = '9\_pibpc\_ppa\_log\_1950.csv'

# Para la string 'TIGRES ASIATICOS' generar un codigo nuevo (a revison para ser desambiguado con posibles matches)

# PRUEBA SOBRE OTRO CSV DE UN TOPICO ARGENDATA CON PROBLEMA:

# TOPICO = 'POBREZ'

# ARCHIVO = 'ISA\_global-poverty\_it3.csv'

# no se registra ni columna similar a geocodigo ni desc\_fundar, a pesar de tener una columna llamada region porque el contenido es ‘GBA' y 'national'

# que son georreferencial pero no son parte de desc\_fundar, y no tiene una columna de geocodigo tampoco

# PRUEBA SOBRE OTRO CSV DE UN TOPICO ARGENDATA CON PROBLEMA:

# TOPICO = 'PRECIO'

# ARCHIVO = '7\_comparacion\_inflacion\_mediana\_argentina\_latam\_1992\_2022.csv'

# No se cumple condición X% Los valores de esta columna son: 'Argentina', 'América Latina y Caribe'

# 'América Latina y Caribe' no aparece en el geonomenclador, si aparece ‘América Latina y el Caribe’.

# No se lee columna similar ni a geocodigo ni a desc\_fundar

# PRUEBA SOBRE OTRO CSV DE UN TOPICO ARGENDATA CON PROBLEMA:

# TOPICO = 'TRANEN'

# ARCHIVO = 'potencia\_instalada\_renov\_regional.csv'

# No se cumple condición X% Los valores de esta columna son: 'CABA y Provincia de Buenos Aires', 'Centro', 'Comahue',

# 'Cuyo', 'Litoral', 'Noreste Argentino', 'Noroeste Argentino', 'Patagonia', 'Total'

# 'CABA y Provincia de Buenos Aires', 'Noreste Argentino', 'Noroeste Argentino' no aparece en el geonomenclador.

# No se lee columna similar ni a geocodigo ni a desc\_fundar

* **Solo considera español desc\_fundar**
* **Csv sin deteccion de columna geocodigo ni deteccion columna desc\_fundar:  
  Hay columnas georreferenciales no detectadas (ejemplo: TOPICO = 'POBREZ', ARCHIVO = 'ISA\_global-poverty\_it3.csv')**
* **No resuelve si un nombre desc\_fundar puede tener más de un código**
* **Solo considera desc\_fundar, por ahora no incluyo busqueda en name\_long**

**Hasta 28/05/2025:**

**15/05/2025**

- pruebo con 60% en lugar de 80%

- queda poner para todos los tópicos en produccion (generacion de carpetas nuevas, los csvs, etc)

- registrar problemas que quedaron dando vueltas (del excel y del procesador de texto)

- categorizar problemas reunion 29, ir categorizando problemas

¿"si un dataset tiene un código que esté incluido en los geocodigos del geonomenclador, se nomencla con la string de name\_long" por "si un dataset tiene un código que esté incluido en los geocodigos del geonomenclador, se nomencla con la string de name\_long (si es que existe), y si no es así se nomencla con la string de desc\_fundar correspondiente"? me dieron el ok asi que las tareas quedan:

- si un dataset tiene un código que esté incluido en los geocodigos del geonomenclador, se nomencla con la string de name\_long (si es que existe), y si no es así se nomencla con la string de desc\_fundar correspondiente

- Si un dataset tiene un código que NO este incluido en los geocodigos y tiene una string que matchee sin ambigüedad en desc\_fundar, entonces se usa el código que está en el geonomenclador (y se guarda la asociación del codigo encontrado con el codigo usado).

- Si un dataset tiene un código que NO este incluido en los geocodigos y tiene una string que NO matchea sin ambigüedad en desc\_fundar, se genera un código nuevo (a revisión para ser desambiguado con posibles matches)

- Si un dataset tiene una string que matchee sin ambigüedad en desc\_fundar entonces se usa el código que está en el geonomenclador para esa string.

- Si un dataset tiene una string que NO matchee sin ambigüedad en desc\_fundar, entonces se genera un código nuevo (a revisión para ser desambiguado con posibles matches)

**05/05/2025**

Cómo seguir:

Quedaron las siguientes definiciones para las imputaciones:

- Si un dataset tiene un código que esté incluido en los geocodigos del geonomenclador, se nomencla con la string de name\_long

- Si un dataset tiene un código que NO este incluido en los geocodigos y tiene una string que matchee sin ambigüedad en desc\_fundar, entonces se usa el código que está en el geonomenclador (y se guarda la asociación del codigo encontrado con el codigo usado).

- Si un dataset tiene un código que NO este incluido en los geocodigos y tiene una string que NO matchea sin ambigüedad en desc\_fundar, se genera un código nuevo (a revisión para ser desambiguado con posibles matches)

- Si un dataset tiene una string que matchee sin ambigüedad en desc\_fundar entonces se usa el código que está en el geonomenclador para esa string.

- Si un dataset tiene una string que NO matchee sin ambigüedad en desc\_fundar, entonces se genera un código nuevo (a revisión para ser desambiguado con posibles matches)

voy a usar como input de geonomeclador: geonomenclador\_LuloTest.csv (viene a ser la pestaña front de https://docs.google.com/spreadsheets/d/1fAQ3rmE3Fr6ztUKXvDYJw0iGH-Vidrx8wYkkHEgBBIQ/edit?gid=1603974855#gid=1603974855)

Tema a resolver:

la pestaña front del excel geonomenclador tiene 4 columnas que son: geocodigo | desc\_fundar | name\_long | name\_short

Pensaba usar esa pestaña para el renaming pero, si no estoy pifiando, no estaría bien porque la columna name\_long no está completa. Por ejemplo, el geocodigo GUY se corresponde al desc\_fundar Guyana, pero su respectiva fila de la columna name\_long está vacia (adjunto captura).

Esto lo vi básicamente porque cuando estaba trabajando en el bullet "si un dataset tiene un código que esté incluido en los geocodigos del geonomenclador, se nomencla con la string de name\_long", casos como este, el name\_long del renaming quedaba vacio.

La pestaña export tiene las columnas:

geocodigo | name\_long | name\_short | iso\_2

Y en ese caso el geocodigo GUY se corresponde al name\_long Guyana

Entonces, cambio el bullet ¿"si un dataset tiene un código que esté incluido en los geocodigos del geonomenclador, se nomencla con la string de name\_long" por "si un dataset tiene un código que esté incluido en los geocodigos del geonomenclador, se nomencla con la string de name\_long (si es que existe), y si no es así se nomencla con la string de desc\_fundar correspondiente"?

**Hasta 05/05/2025:**

Versión 1 renaming

INPUTS:

-'georef\_keys.txt': ['country\_name\_abbreviation', 'pais', 'partner\_name\_short\_en', 'region', 'iso3\_desc', 'partner\_region\_name\_short\_es', 'region\_code', 'partner\_region\_id', 'pais\_o\_region', 'pais\_o\_grupo\_de\_paises', 'iso3', 'codigo\_pais', 'country\_code', 'partner\_code', 'iso3c', 'countryname', 'nombre\_pais', 'region\_name', 'iso3\_desc\_fundar', 'pais\_nombre']

- geonomenclador.csv

--------------------

- Ya detectadoa los .csv por tópico de argendata que tengan columnas que se llamen según 'georef\_keys.txt

- Ya detectados los .csv por tópico de argendata que tengan columnas que cumplan: si al menos el 80% de los valores (\*) están contenidos en el conjunto de valores de la columna del geonomenclador (eliminando nulos y homogeneizando temas: espacios "NEA "|"NEA", mayúsculas: "pampeana" | "Pampeana", tildes: "AFGANISTÁN" | "AFGANISTAN" )

- Registro de que tiene el geonomenclandor.csv (según coincidencia 80%) pero no el 'georef\_keys.txt' [df\_GeoNomYes\_GeoKeysNo]

con eso se contribuye a sumar nombres de POSIBLES columnas georreferenciales a 'georef\_keys.txt'

- Registro de que tiene 'georef\_keys.txt' pero no la condición geonomenclandor.csv (según coincidencia 80%) [df\_GeoKeysYes\_GeoNomNo]

- acá me gustaría que haya, hasta el momento, solo NO detección de columnas en idiomas distintos al español:

pero hay otros problemas (archivosArgendata\_problemaRenaming.ods)

- geonomenclador tiene un tema con los geocódigos, un "mismo" name\_lon por tiene distinto geocodigos:

ejemplo "América del Norte" tiene 3 geocódigos distintos...

si algo dice "Europa" solo, y en el geonomenclados hay más de 5 opciones que contienen "Europa"... cuál tomamos?:

porque podría "jugar" con una detección de que si tal palabra está contenida en name\_lon considero esa fila... pero "Europa" está en más de una opción

- si la columna iso3 aparece vamos bien de todas maneras

- tema condicion geonomenclador 80% (detección de .csv con problemas), la bajamos a 60 %? (lo quioero explorar)

- geonomenclador tiene un tema con los geocódigos, por ejemplo "América del Norte" tiene 3 geocódigos distintos

- si algo dice "Europa" solo, y en el geonomenclados hay más de 5 opciones que contienen "Europa"... cuál tomamos?

- tema idioma (detección de .csv con problemas)

- los csv con tema idioma + que muchos no estarán en el geonomenclador los quiero ver en detalle -> a ver si al aplicar tema idioma igual cumple el

- tema encoding, guardado (proceso renaming)

- multiples codigos para un america del norte, dejo registrado que unico codigo para ese string -s hago uno para uno

- si aparece un codigo en el csv, le mando csv

- si no tengo codigo, le asigno

PRODUCTO DE ESTO: v1\_archivosArgendata\_problemaRenaming.ods -> DETALLE DE ARCHIVOS CON GEORREFERENCIA QUE NO SON DETECTADOS Y SU POR QUÉ