Islas Canarias
Del 15 al 19 de noviembre de 2021

istac | INSTITUTO CANARIO DE ESTADÍSTICA

arte_

XXI Jornadas de Estadística de las Comunidades Autónomas

Automatizando la carga de datos en el portal SIEMAC

Sergio Delgado Quintero
Samuel Baute García
Rita Díaz Adán

@sdelquin @samuelbaute @ritadiazadan

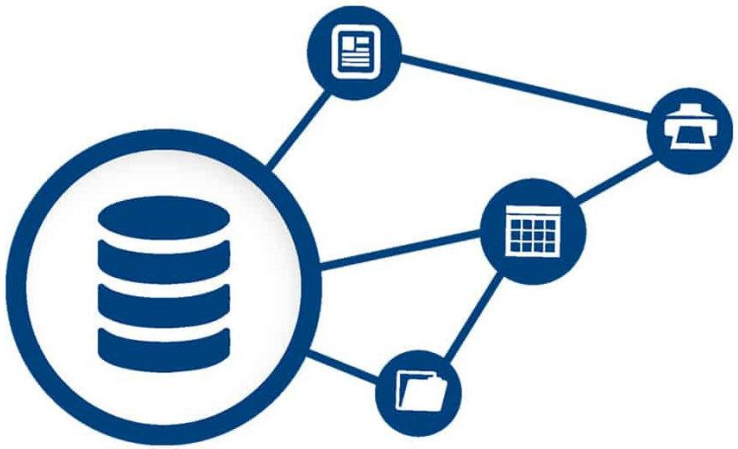
@sdelquin @samuelbaute @rdiaada

Introducción

¿Qué vamos a ver?

- Qué es SIEMAC
- Modernización
- Recolección de datos
- Automatización

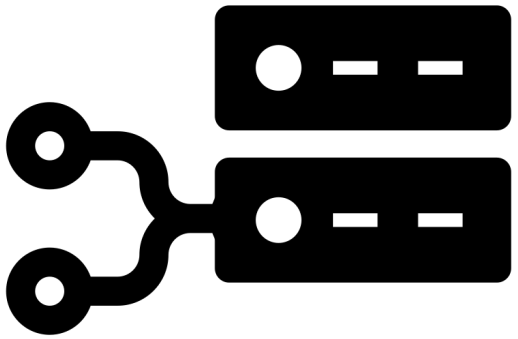
SIE mac



eurostat

SIE-MAC

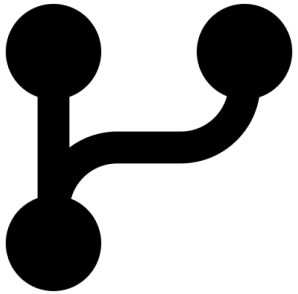
- Proyecto INTERREG III-B (Azores-Madeira-Canarias)
- Sistema de Información Estadística de la Macaronesia
- Base de datos común



Modernización de la plataforma

Plan de modernización

- Web y CMS obsoletos
- Datos desactualizados
- Rehacer > Mantener



Nuevo CMS

wpDataTables

WPML

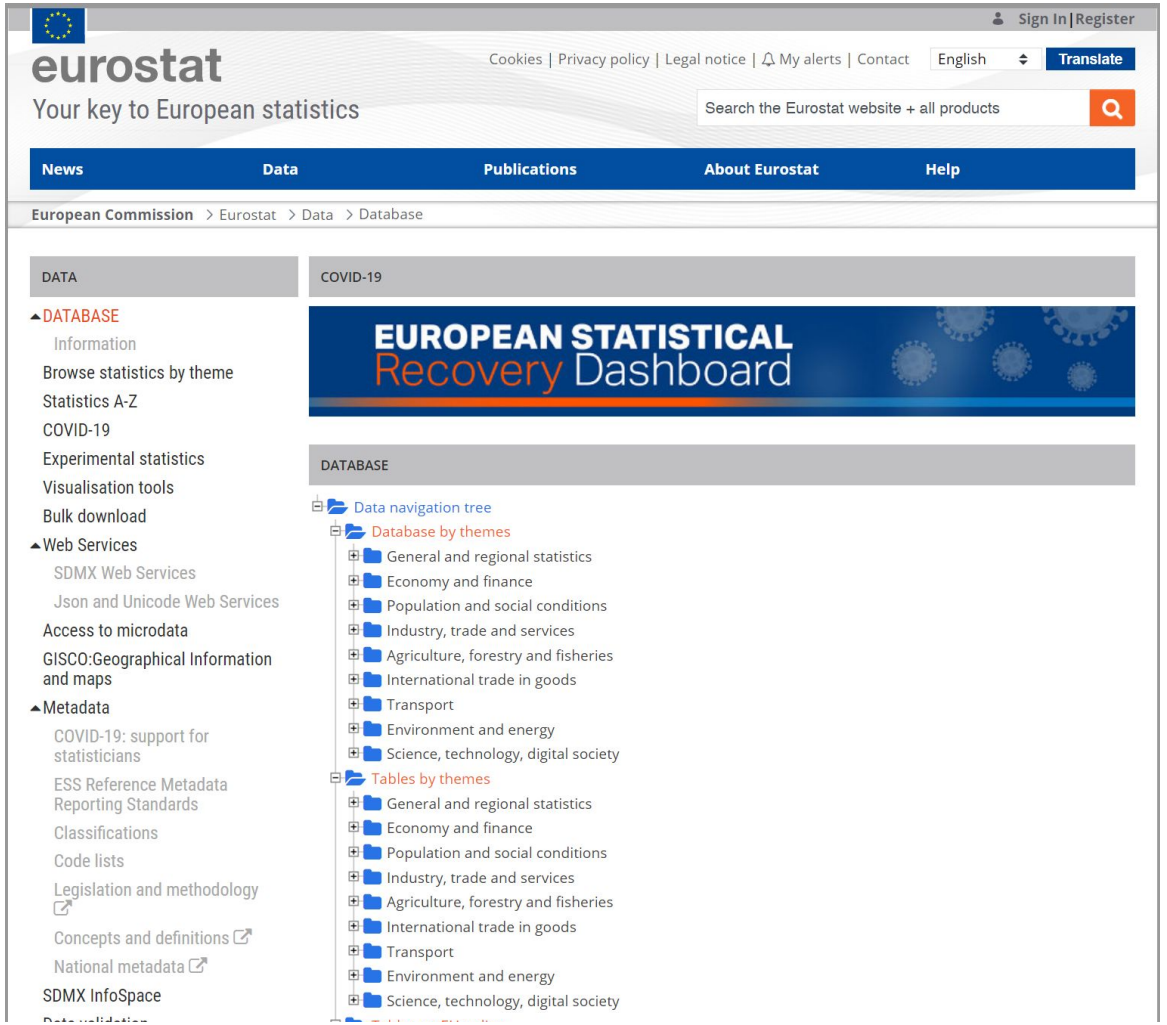
¿Y los datos?

Recolección y normalización de datos

Unificar fuentes de datos < Aprovechar datos abiertos

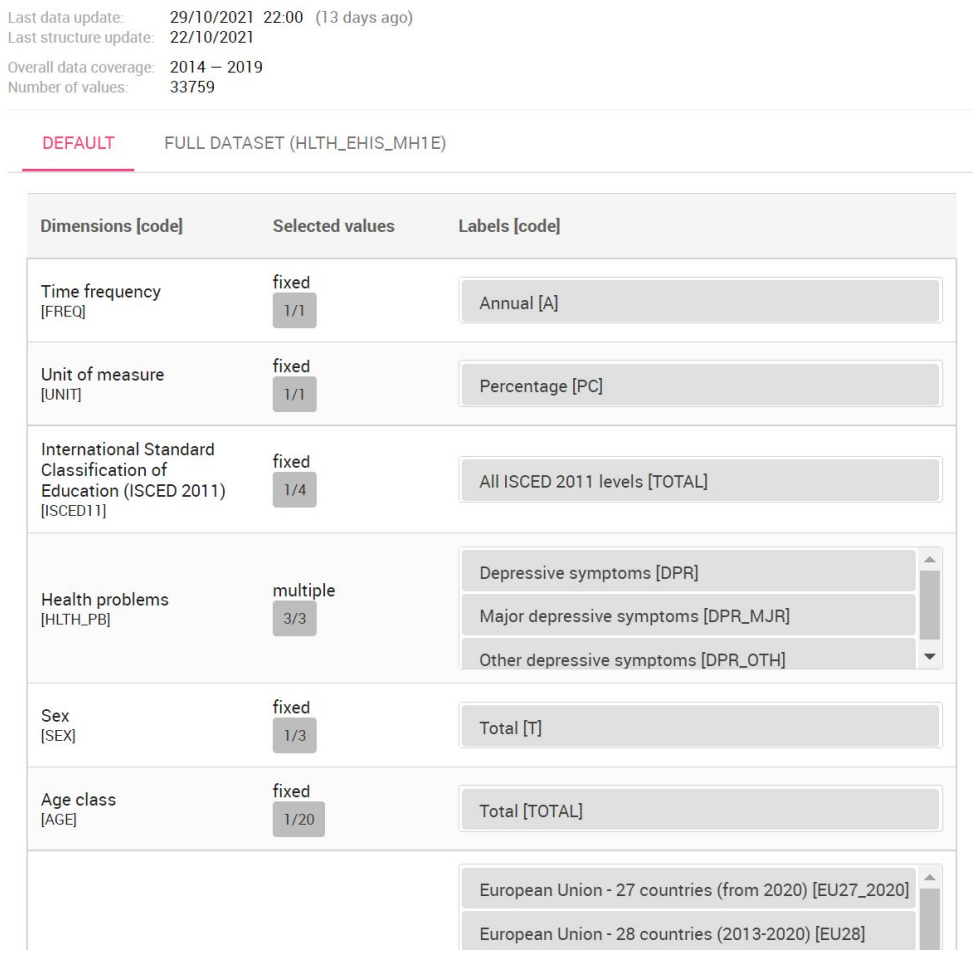
Eurostat

- Ya se incluyen los datos por NUTS
- Tenemos ya los datos limpios y armonizados
- Exportable en formatos abiertos



Problemas

- Activos semánticos y datos solo en inglés
- Clasificaciones sin migrar aún a la Infraestructura de Datos y Metadatos Estadísticos del ISTAC (eDatos)
- El plugin solo admite formato JSON, y Eurostat exporta en TSV



Last data update: 29/10/2021 22:00 (13 days ago)
Last structure update: 22/10/2021
Overall data coverage: 2014 – 2019
Number of values: 33759

DEFAULT | FULL DATASET (HLTH_EHIS_MH1E)

Dimensions [code]	Selected values	Labels [code]
Time frequency [FREQ]	fixed 1/1	Annual [A]
Unit of measure [UNIT]	fixed 1/1	Percentage [PC]
International Standard Classification of Education (ISCED 2011) [ISCED11]	fixed 1/4	All ISCED 2011 levels [TOTAL]
Health problems [HLTH_PB]	multiple 3/3	Depressive symptoms [DPR] Major depressive symptoms [DPR_MJR] Other depressive symptoms [DPR_OTH]
Sex [SEX]	fixed 1/3	Total [T]
Age class [AGE]	fixed 1/20	Total [TOTAL]
		European Union - 27 countries (from 2020) [EU27_2020] European Union - 28 countries (2013-2020) [EU28]

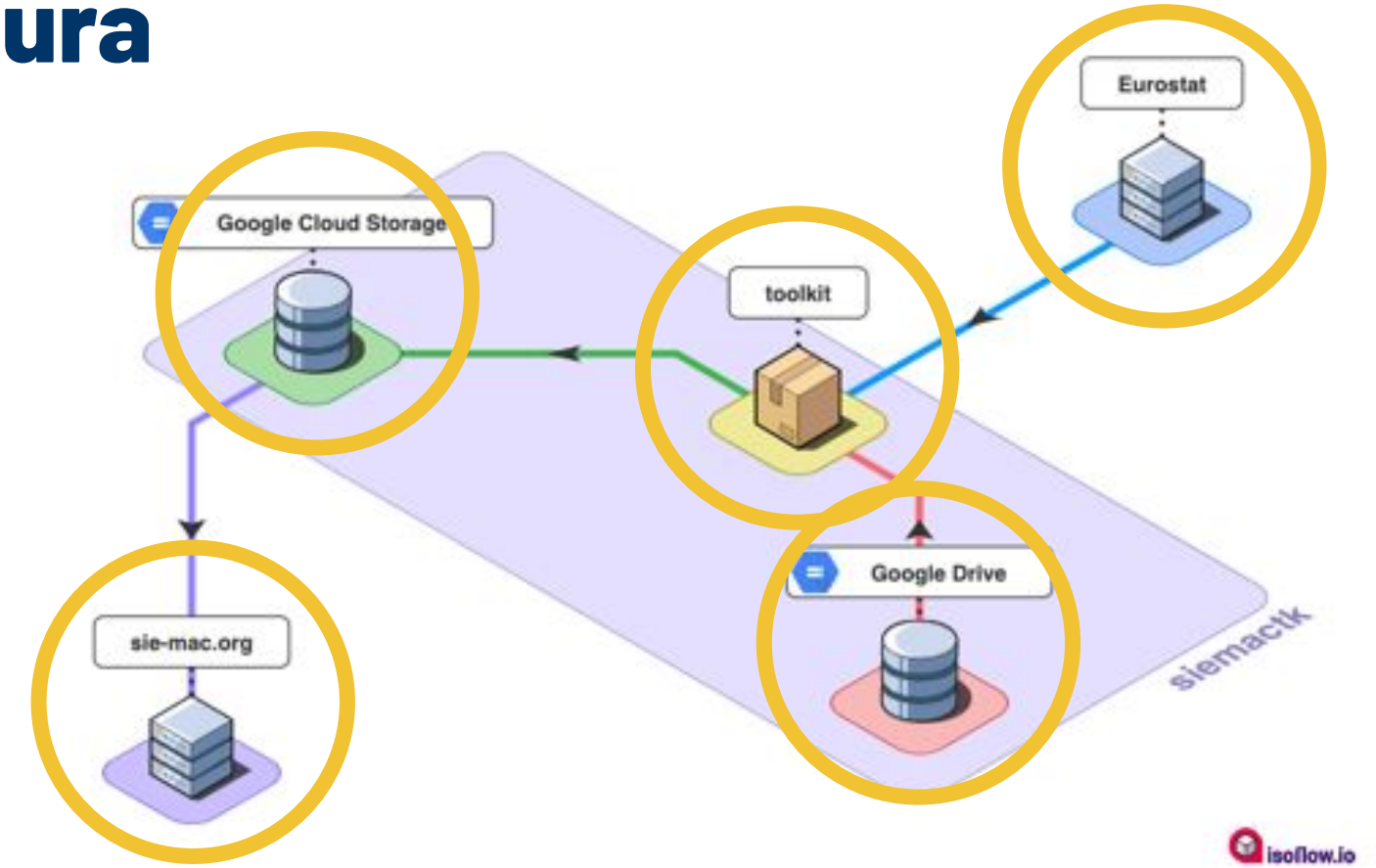
Soluciones

- Traducir todos los activos semánticos -> 1.257 códigos de 46 CL
- Necesario un proceso que realice las traducciones y la conversión TSV a JSON

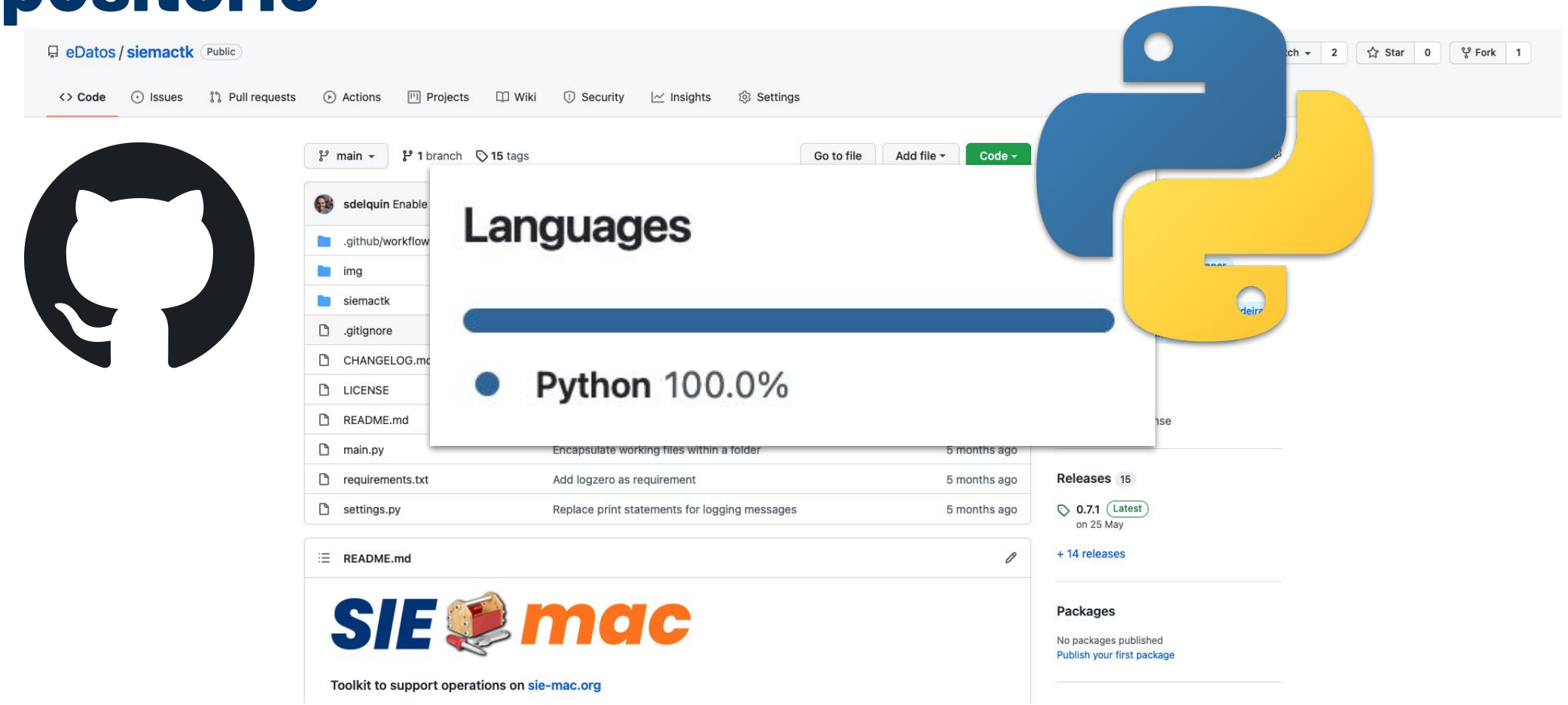
	A	B	C	
1	cl	code	ES	PT
1245	victim	INJ	Heridos	Lesionados
1246	worktime	TOTAL	Total	Total
1247	worktime	PT	A tiempo parcial	Part-time
1248	worktime	FT	Tiempo completo	A tempo inteiro
1249	worktime	NRP	Sin respuesta	Sem resposta
1250	wstatus	EMP	Personas empleadas	Pessoas empregadas
1251	wstatus	SAL	Empleados	Empregados
1252	wstatus	NSAL	Trabajadores por cuenta ajena, excepto los asalariados	Pessoas empregadas excepto empreg
1253	wstatus	POP	Población	População
1254	wstatus	ACT	Personas activas	Pessoas activas
1255	wstatus	UNE	Desempleados	Desempregados
1256	wstatus	INAC	Inactivos	Pessoas inactivas
1257	wstatus	UNK	Desconocido	Desconhecido
1258	wstatus	NRP	Sin respuesta	Sem resposta
1259				

Manipulación de datos y automatización

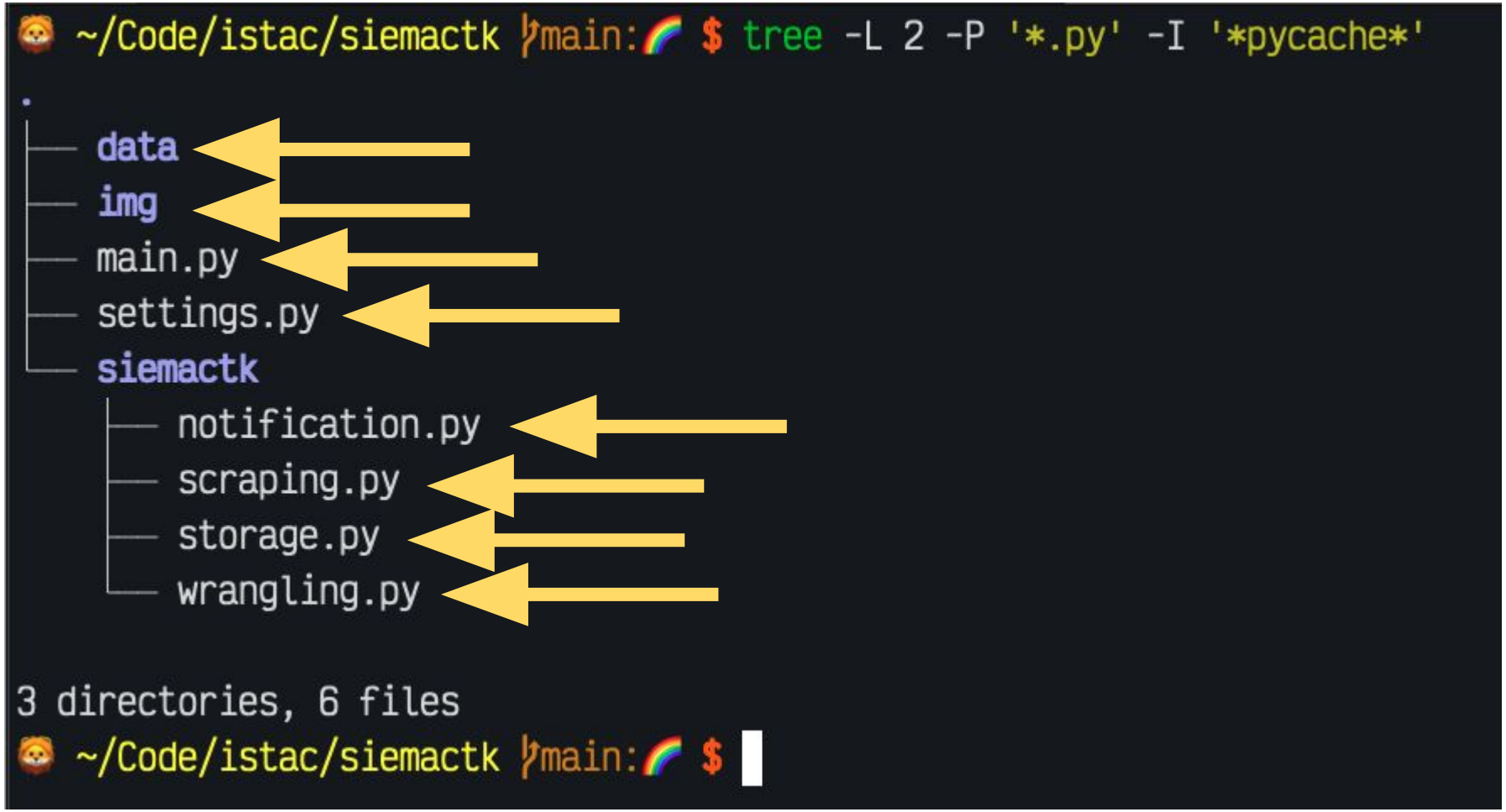
Arquitectura



Repositorio



Código


~/Code/istac/siemactk main: $ tree -L 2 -P '*.py' -I '*pycache*'
.
data
img
main.py
settings.py
siemactk
notification.py
scraping.py
storage.py
wrangling.py
3 directories, 6 files
~/Code/istac/siemactk main: $

Scraping

yagdrive

https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?file=data/agr_r_accts.tsv.gz


```
import gzip
```

```
agr_r_accts.tsv
```

Tratamiento (i)

	indic_ag,itm_newa,unit,geo\time	2019	2018	2017	2016	2015	2014	2013	2012	2011	...	1982	1981	1980	1979	1978
0	PROD_BP,01000,MIO_EUR,AT	800.27	776.49	754.07	749.33	728.38	769.41	717.58	1029.21	906.72	...	:	:	:	:	:
1	PROD_BP,01000,MIO_EUR,AT1	463.42	446.07	403.65	444.80	413.60	449.88	428.48	519.63	516.20	...	:	:	:	:	:
2	PROD_BP,01000,MIO_EUR,AT11	87.50	88.60	82.87	89.85	80.75	86.74	71.76	115.35	90.22	...	:	:	:	:	:
3	PROD_BP,01000,MIO_EUR,AT12	373.81	355.54	319.25	352.70	330.98	361.45	354.97	402.61	423.85	...	:	:	:	:	:
4	PROD_BP,01000,MIO_EUR,AT13	2.10	1.93	1.53	2.24	1.88	1.70	1.76	1.68	2.12	...	:	:	:	:	:
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
285373	TAX,37000,MIO_NAC,UKM7	:	: z	:	:	:	:	:	:	:	...	:	:	:	:	:
285374	TAX,37000,MIO_NAC,UKM8	:	: z	:	:	:	:	:	:	:	...	:	:	:	:	:
285375	TAX,37000,MIO_NAC,UKM9	:	: z	:	:	:	:	:	:	:	...	:	:	:	:	:
285376	TAX,37000,MIO_NAC,UKN	:	: z	:	:	:	:	:	:	:	...	:	:	:	:	:
285377	TAX,37000,MIO_NAC,UKN0	:	: z	:	:	:	:	:	:	:	...	:	:	:	:	:

285378 rows × 48 columns

pandas

1. Filtrado por códigos geográficos
2. Limpieza de valores nulos
3. Reemplazo de punto flotante
4. Recodificación a los idiomas ES y PT
5. Exportación a json y tsv

`agr_r_accts.tsv`




Tratamiento (ii)

	indic_ag	itm_newa	unit	geo	2019	2018	2017	2016	2015	2014
0	Valor de la producción a precio básico	Cereales (incluidas las semillas)	Millones de euros	Canarias	NaN	0.54	0.53	3.18	0.56	0.56
1	Valor de la producción a precio básico	Cereales (incluidas las semillas)	Millones de unidades de la moneda nacional	Canarias	NaN	0.54	0.53	3.18	0.56	0.56
2	Valor de la producción a precio básico	Trigo y escanda	Millones de euros	Canarias	NaN	0.07	0.07	0.34	0.06	0.06
3	Valor de la producción a precio básico	Trigo y escanda	Millones de unidades de la moneda nacional	Canarias	NaN	0.07	0.07	0.34	0.06	0.06
4	Valor de la producción a precio básico	Trigo blando y escanda	Millones de euros	Canarias	NaN	0.07	0.07	0.34	0.06	0.06

🇪🇸 `agr_r_accts_es.tsv`

	indic_ag	itm_newa	unit	geo	2019	2018	2017	2016	2015	2014
0	Valor de produção a preço de base	Cereais (incluindo sementes)	Milhões de euros	Canárias	NaN	0.54	0.53	3.18	0.56	0.56
1	Valor de produção a preço de base	Cereais (incluindo sementes)	Milhões de unidades monetárias nacionais	Canárias	NaN	0.54	0.53	3.18	0.56	0.56
2	Valor de produção a preço de base	Trigo e espelta	Milhões de euros	Canárias	NaN	0.07	0.07	0.34	0.06	0.06
3	Valor de produção a preço de base	Trigo e espelta	Milhões de unidades monetárias nacionais	Canárias	NaN	0.07	0.07	0.34	0.06	0.06
4	Valor de produção a preço de base	Trigo mole e espelta	Milhões de euros	Canárias	NaN	0.07	0.07	0.34	0.06	0.06

🇵🇹 `agr_r_accts_pt.tsv`

`{} df.to_json()`

Almacenamiento

Detalles del bucket

Nombre	Tamaño	Tipo	Fecha de creación	Clase de almacenamiento	Última modificación
agr_r_accts_es.json	1.3 MB	application/json	1 oct. 2021 08:06:21	Standard	1 oct. 2021 08:0...
agr_r_accts_es.tsv	392.5 KB	text/tab-separated-values	1 oct. 2021 08:06:20	Standard	1 oct. 2021 08:0...
agr_r_accts_pt.json	1.3 MB	application/json	1 oct. 2021 08:06:22	Standard	1 oct. 2021 08:0...
agr_r_accts_pt.tsv	385.1 KB	text/tab-separated-values	1 oct. 2021 08:06:21	Standard	1 oct. 2021 08:0...
agr_r_animal_es.json	91 KB	application/json	1 oct. 2021 08:06:24	Standard	1 oct. 2021 08:0...
agr_r_animal_es.tsv	32.1 KB	text/tab-separated-values	1 oct. 2021 08:06:24	Standard	1 oct. 2021 08:0...
			1 08:06:25	Standard	1 oct. 2021 08:0...
			1 08:06:24	Standard	1 oct. 2021 08:0...
			1 08:09:24	Standard	1 oct. 2021 08:0...
			1 08:09:24	Standard	1 oct. 2021 08:0...
bd_hgnace2_r3_pt.json	336.6 KB	application/json	1 oct. 2021 08:09:25	Standard	1 oct. 2021 08:0...
bd_hgnace2_r3_pt.tsv	200 KB	text/tab-separated-values	1 oct. 2021 08:09:24	Standard	1 oct. 2021 08:0...
bd_size_r3_es.json	145.5 KB	application/json	1 oct. 2021 08:08:38	Standard	1 oct. 2021 08:0...
bd_size_r3_es.tsv	93.4 KB	text/tab-separated-values	1 oct. 2021 08:08:38	Standard	1 oct. 2021 08:0...
bd_size_r3_pt.json	149.4 KB	application/json			
bd_size_r3_pt.tsv	90.9 KB	text/tab-separated-v			
demo_r_d2jan_es.json	612.3 KB	application/json			
demo_r_d2jan_es.tsv	222.7 KB	text/tab-separated-v			
demo_r_d2jan_pt.json	613.4 KB	application/json			
demo_r_d2jan_pt.tsv	226.1 KB	text/tab-separated-v			
demo_r_d3dens_es.json	2 KB	application/json			
demo_r_d3dens_es.tsv	1 KB	text/tab-separated-v			
demo_r_d3dens_pt.json	2 KB	application/json			
demo_r_d3dens_pt.tsv	1 KB	text/tab-separated-v			
demo_r_find2_es.json	4.4 KB	application/json			
demo_r_find2_es.tsv	1.9 KB	text/tab-separated-v			
demo_r_find2_pt.json	4.4 KB	application/json			

Filas por página: 50 1 – 50 de 232

```
from google.cloud import storage

gcs = storage.Client()
bucket = gcs.get_bucket('siemac')
```

Notificación

📊 Actualización de datos SIEMAC

informatica.istac@gobiernodecanarias.org
para mí, edatos

Hola,

Se ha realizado la subida de datos SIEMAC tras el scraping a EUROSTAT y su posterior procesamiento.

Relación de ficheros con sus URLs de descarga:

agr_r_accts
agr_r_accts_es.tsv
agr_r_accts_es.json
agr_r_accts_pt.tsv
agr_r_accts_pt.json

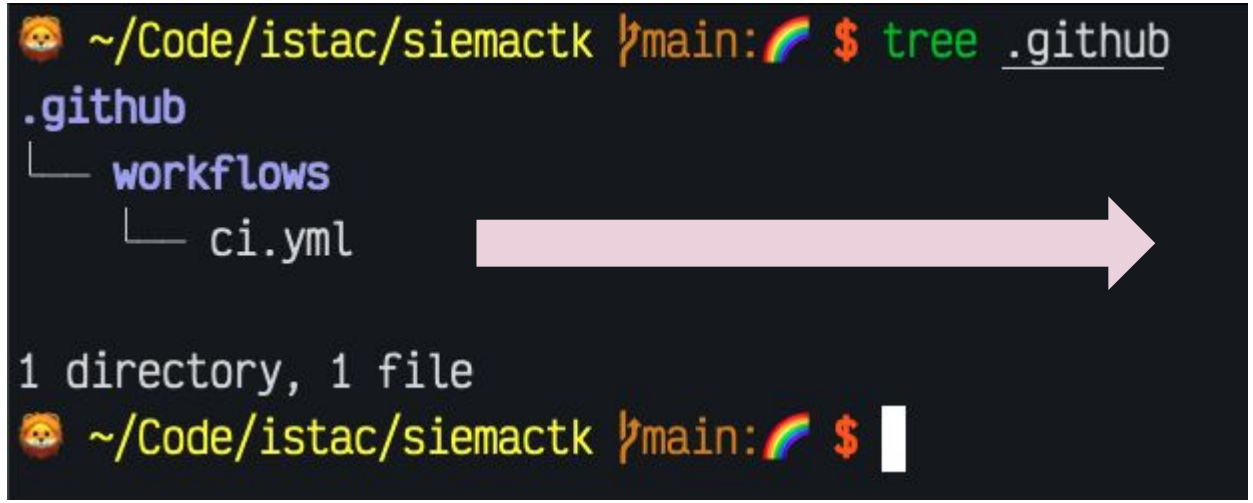
agr_r_animal
agr_r_animal_es.tsv
agr_r_animal_es.json
agr_r_animal_pt.tsv
agr_r_animal_pt.json

demo_r_d2jan
demo_r_d2jan_es.tsv
demo_r_d2jan_es.json
demo_r_d2jan_pt.tsv
demo_r_d2jan_pt.json

demo_r_d3dens
demo_r_d3dens_es.tsv
demo_r_d3dens_es.json
demo_r_d3dens_pt.tsv
demo_r_d3dens_pt.json

demo_r_pjanind2
demo_r_pjanind2_es.tsv
demo_r_pjanind2_es.json
demo_r_pjanind2_pt.tsv
demo_r_pjanind2_pt.json

Automatización



37 lines (37 sloc) | 1.32 KB

```yaml
name: CI
on:
  schedule:
    - cron: "0 7 1 * *"
  workflow_dispatch:
jobs:
  eurostat-scraping:
    name: Scrap Eurostat datasets for sie-mac.org
    runs-on: ubuntu-latest
```

GitHub Actions

~ 7 minutos

Propuestas de mejora

- Mejorar el control de errores.
- Optimizar los procesos de tratamiento de datos.
- Estudiar integración con API de wordpress.

Islas Canarias
Del 15 al 19 de noviembre de 2021

¡Gracias!

istac | INSTITUTO CANARIO DE ESTADÍSTICA

arte_

@sdelquin @samuelbaute @ritadiazadan

@sdelquin @samuelbaute @rdiaada