Providing Online Based Enterprise Characteristics with the Web Intelligence Hub.

Experience of the Web Intelligence Network activities

Jacek Maślankowski (j.maslankowski@stat.gov.pl)
Statistics Poland
University of Gdańsk

2025/02/04

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement numer: 101035829 (2020-PL-SmartStat)

Web Intelligence
Network

Funded by the
European Union

Funded by
the European Union

Agenda

Background information

Methodology and data processing

Main findings and conclusions

Funded by the European Union

What is Online Based Enterprise Characteristics?

OBEC – term have been used since 2018

The use of a website by the enterprise to present its ‘business’, with extension to social media perspective.

It includes not only the existence of a website which is located on servers belonging to the enterprise or at one of the enterprise’s sites, but also third party’s websites.

How can we define OBEC population?

Methodological manual for data compilers and users of the ICT survey:
A4. Does your enterprise have a website?

The population of OBEC use case consists of enterprises having website and employing 10 or more employees.

Funded by the European Union

OBEC population = URL database for WIH

It consists of two datasets:

- For the Web Intelligence Platform (web scraping tool with data storage Amazon OpenSearch):
  - Anonymized ID
  - Enterprise URL
  - Group to which it belongs (e.g. /OBEC)
- For further processing in Datalab (JupyterLab with Python, RStudio):
  - Business register ID
  - Anonymized ID
  - Other attributes

To link with business register

Funded by the European Union

What indicators OBEC can provide?

- Main indicators
  - Social Media Presence
    - Type of social media
    - Purpose of social media
  - E-commerce
    - E-shops
    - Product list
- Other indicators
  - User friendliness
    - GDPR policy
    - Accessibility
    - Sustainability
  - Climate neutrality
    - Green economy
    - SDG support
  - Innovations
    - Patents
    - Collaboration with Universities
    - Address - location - technical incubator
    - Innovative companies and regions

Web Intelligence Network

Funded by the European Union

List of indicators

INDICATOR	BREAKDOWN1	BREAKDOWN2	…	MEASURE
Social media presence	Instagram/X/Facebook/LinkedIn/Xing/Tiktok/ Pinterest/?Viadeo/?Yammer/Flickr/Instagram/Snapchat/ Slideshare/Threads/Whatsapp/Reddit/Telegram/Discord/WeChat/	NACE	Company size, Region	Present / Not in any; in specific
E-commerce	-	NACE	Company size, Region	Yes / No
Chatbot	-	NACE	Company size, Region	Yes / No
Innovation	With extension to other keywords	NACE	Company size, Region	Yes / No
Multilanguage support	-	NACE	Company size, Region	Yes / No
Extracting contact information	Phone number, e-mail address, zip code	NACE	Company size, Region	Yes / No

What is the reason to include enterprises having 10+ employees?

Traditional questionnaire – Survey on ICT Usage and E-commerce in Enterprises

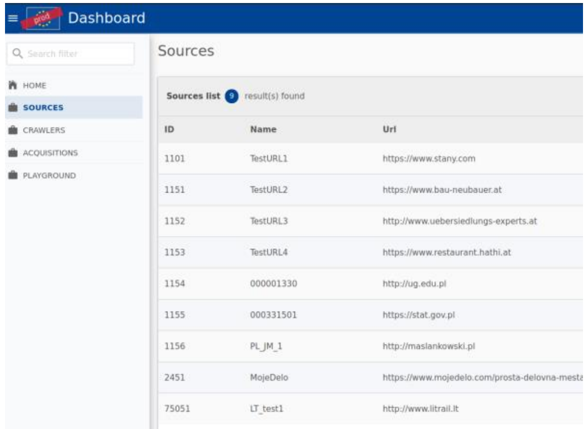
URL database (Uniform Resource Locator – website address)

Significant percentage of small enterprises does not have a website compared to larger enterprises

Names of small companies are usually not unique

Funded by the European Union

WIP for the OBEC



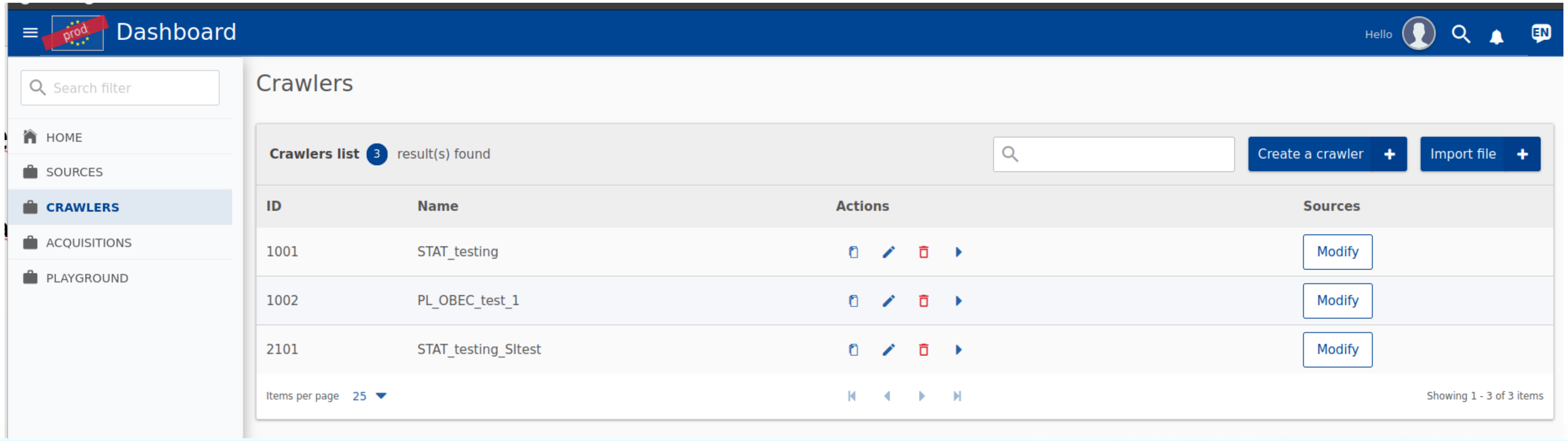
Sources

Sources list 9 result(s) found

ID	Name	Url
1101	TestURL1	https://www.stany.com
1151	TestURL2	https://www.bau-neubauer.at
1152	TestURL3	http://www.uebersiedlungs-experts.at
1153	TestURL4	https://www.restaurant.hathi.at
1154	000001330	http://ug.edu.pl
1155	000331501	https://stat.gov.pl
1156	PL_JM_1	http://maslankowski.pl
2451	MojeDelo	https://www.mojedelo.com/prosta-delovna-mesta
75051	LT_test1	http://www.litrail.lt

General rules:

- one unified data storage for all countries
- each country is contributing to the list of URLs by managing and maintaining
- websites are downloaded for further data processing
- scalable – data stored in NoSQL database



Crawlers

Crawlers list 3 result(s) found

ID	Name	Actions	Sources
1001	STAT_testing		Modify
1002	PL_OBEC_test_1		Modify
2101	STAT_testing_SItest		Modify

Items per page 25

Showing 1 - 3 of 3 items

Funded by the European Union

Prerequisites for OBEC.

What we need to create statistics with WIH?

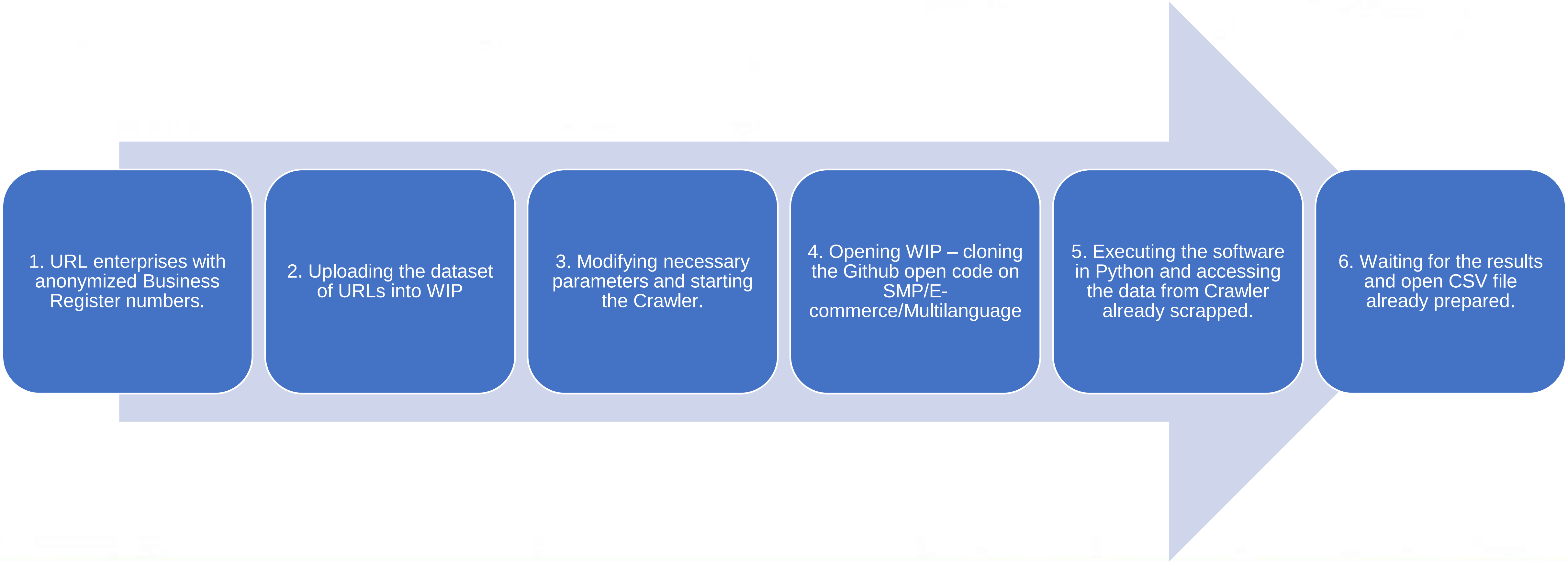
- Access
  - WIN WIP (https://prod.wihp.ecdp.tech.ec.europa.eu/screen/home)
  - WIN Datalab (https://prod.wihp.ecdp.tech.ec.europa.eu/screen/home)

- Scripts
  - Latest release on GitHub
  - JSON file with URLs for web scraping

- Skills (basics)
  - Python
  - Amazon OpenSearch
  - Linux

Funded by the European Union

Collecting and processing data in 6 steps



1. URL enterprises with anonymized Business Register numbers.

2. Uploading the dataset of URLs into WIP

3. Modifying necessary parameters and starting the Crawler.

4. Opening WIP – cloning the Github open code on SMP/E-commerce/Multilanguage

5. Executing the software in Python and accessing the data from Crawler already scrapped.

6. Waiting for the results and open CSV file already prepared.

Funded by the European Union

Methodological obstacles identified

OBEC

- Target population may be unknown
- Providing the list of URLs may be difficult for selected ESS countries
- Different rules apply to websites in different countries, e.g. for some countries it is obligatory to provide tax number on the website while for others not

Funded by the European Union

Other issues

Legal aspects

Web scraping policy

Changes in rules of sharing URL population and data by NSIs

Main findings

- Varying legal regulations at national level in terms of the transfer of data from registers
- The inconsistency of the algorithms for data collection and processing
- The discrepancy between the data obtained with the use of different methods are among the issues under particular consideration within the TSS-WIN project

Conclusions

Adaptation of ESSnet Big Data I, II, Eurostat work

Strict co-operation with Eurostat

Key success factor – the data on OBEC disseminated in the Eurostat Database

Thank you!

Jacek Maślankowski, j.maslankowski@stat.gov.pl (Statistics Poland)

Funded by the European Union