From web to data: Selective scraping & WIN. The Hackathon

Olav ten Bosch, Statistics Netherlands

NTTS 2025, March, Brussels

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Web Intelligence Network

Funded by the European Union

Contents

- Web data in official statistics, why?
- Experiences Statistics Netherlands
- Experiences from the WIN project, WP3 on new use cases
- Bulk scraping versus selective / statistical scraping
- WIN, the hackathon

Why web data in official statistics?

Administrative sources

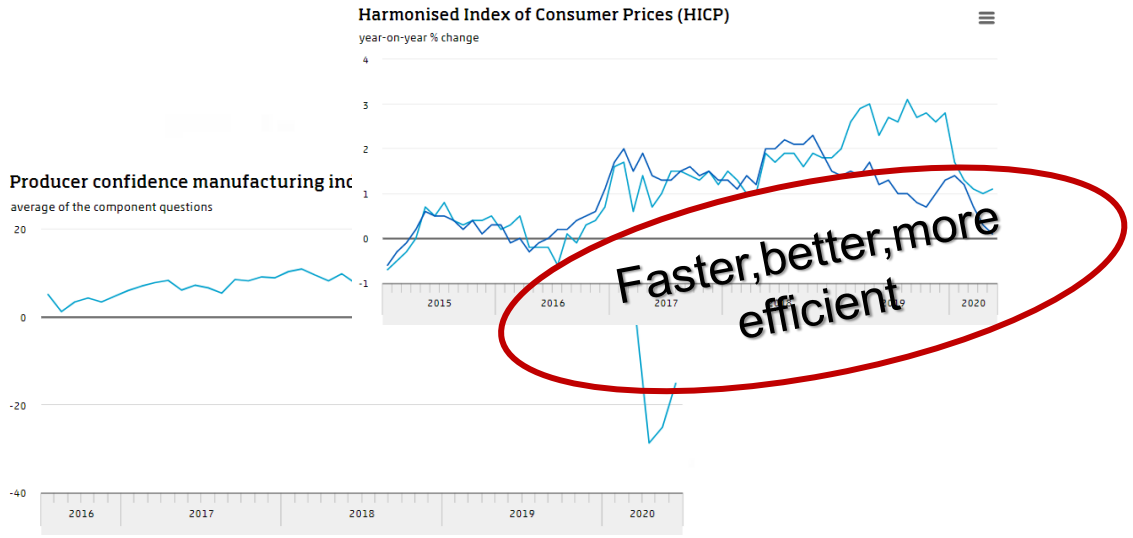
- Tax, social security
- Municipalities/ Provinces
- Supermarkets
- …

Enhancing registers

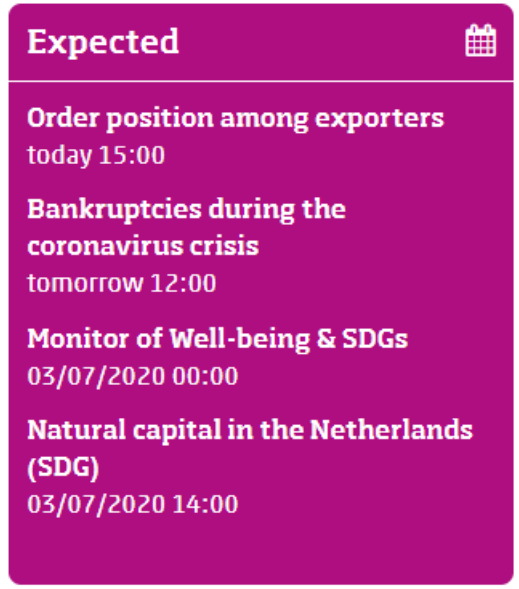
Statistical population discovery

Internet sources

- …
- Surveys *Less!!!*



New indicators



Web Intelligence Network

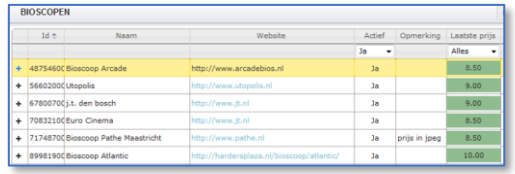
Funded by the European Union

>15 years of web data at Statistics Netherlands

2008-2010
Fuel prices
Real estate
Airtickets

2011-2019
Experimenting
towards offstats

- ***Webshops***: CPI (inflation): prices (clothing), books, travel, consumer electronics
- ***Enterprise websites***: ecommerce, webshop detection, social media use, NACE (SBI), innovative companies, family businesses, drone companies, use of internet standards, platform economy
- ***Annual reports***: financial and institutional data
- ***Social media***: social tension indicator, (social) networks, community statistics
- ***Property portals***: housing market dynamics
- ***Job portals***: trends on job market, skills
- ***Hotels / holiday homes portals***: tourism
- ***Wikipedia***: community data, i.e. on international enterprises, network topology of train tracks, ..
- ***DNS***: domain dynamics / relation with organisations
- ***Municipality portals***: environmental permits
- ***School portals***: courses offered; education trends
- ***Opengov data:*** rental disputes



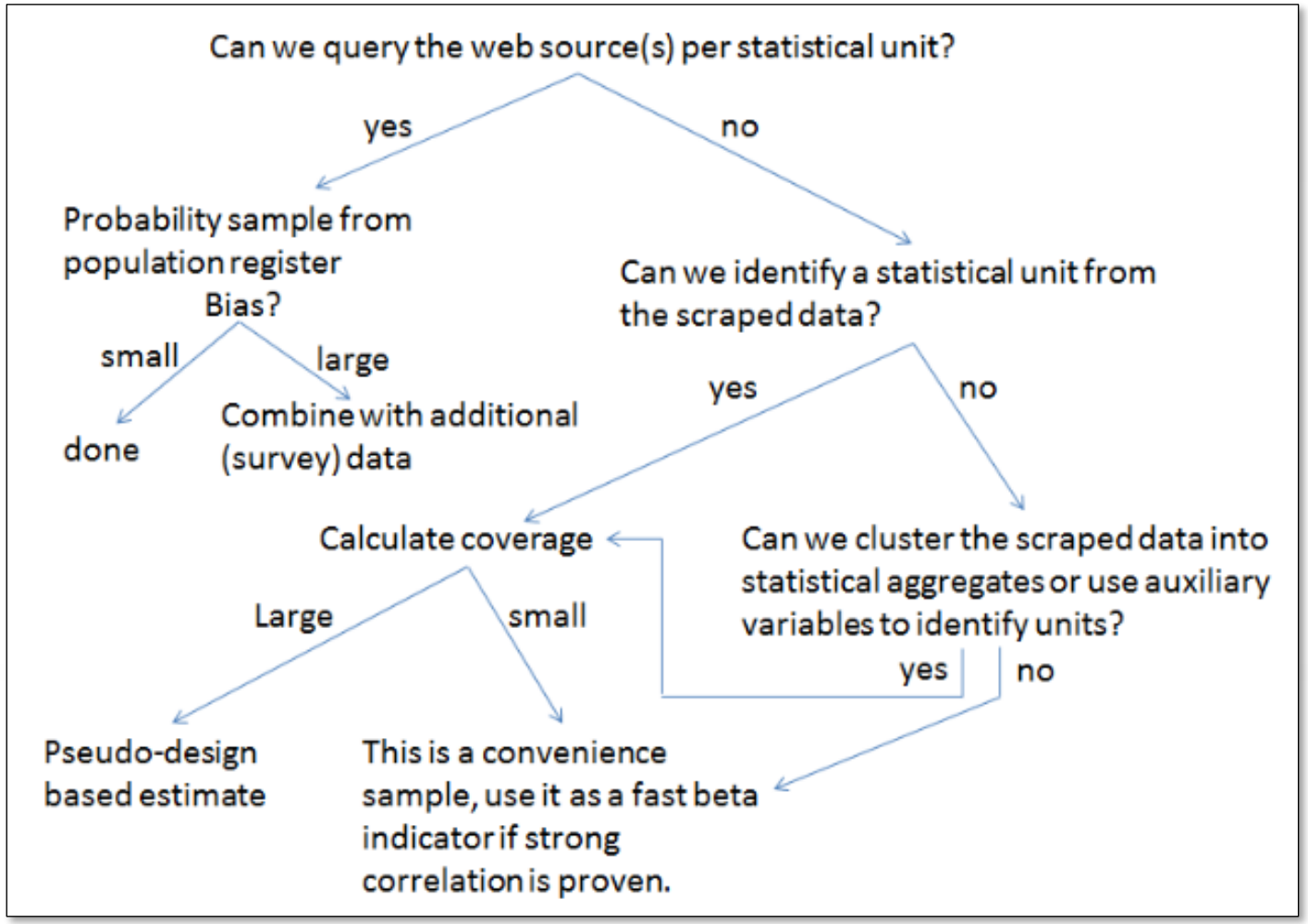
Semi-automatic price collection 2012 ->

2020

web data should be combined

or gathered using a statistical design

Generic workflow for webdata



Bigsurv, 2018

https://www.researchgate.net/publication/327385487_Web_scraping_meets_survey_design_combining_forces

WIN WP3 on new use cases

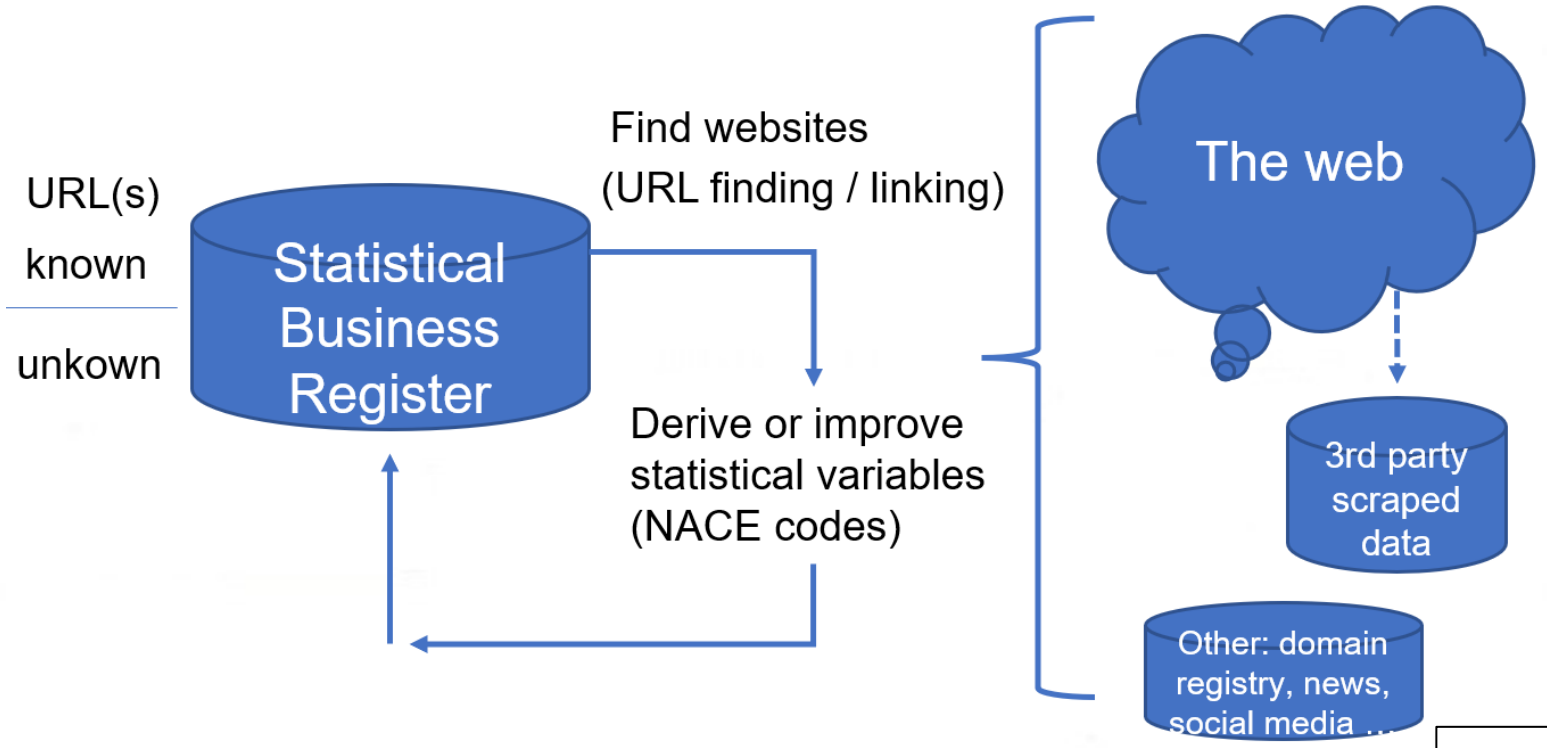
- Exploration of ‘new’ web data sources for the production of official statistics, as primary or auxiliary datasource
- 6 use cases (UCs):

UC1 Characteristics of the real estate market	**PL**, BG, DE-HSL/BBB, FI, FR
UC2 Construction activities	**DE-HSL**, DE-BBB, SE
UC3 Online prices of household appliances and audio-visual, photographic and information processing equipment (and generalising the data collection to other activities)	**SE,** BG
UC4 Experimental indices in tourism statistics (hotel prices)	**PL,** BG
UC5 Business register quality enhancement	**NL,** AT, DE-HSL, SE, FI
UC6 Faster Economic Indicators using new data sources	**SE,** UK

Challenges


Weighting issues
Hedonic prices
Deduplication
UC3: Online Prices
Find the right urls / texts
Missing data
UC2: Construction
UC5: BR quality enhancement
ML models
Experimental studies
Search-engine leakage
UC1: Real Estate
API legals
UC4: Tourism
UC6: Traffic cameras
Stability
Changed privacy concerns
Coverage
Duplicate offers
Lack of standardisation among sites
Lack of data

UC5: business register enhancement

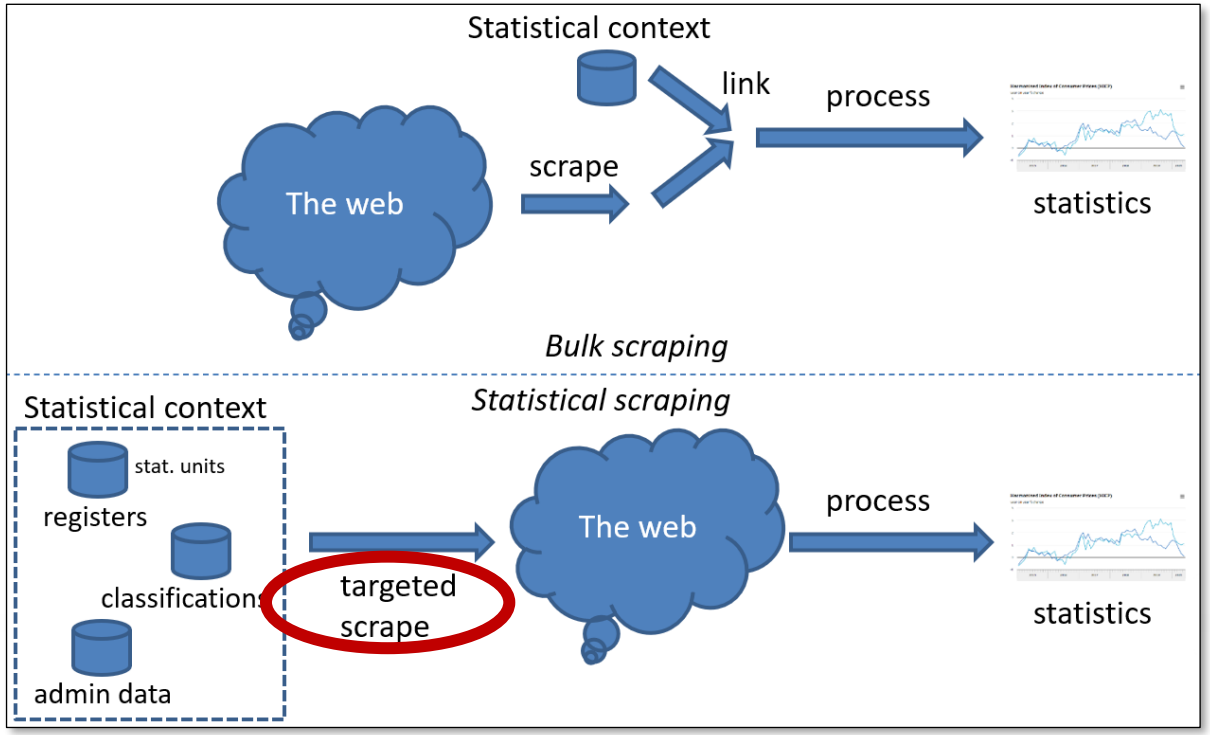


We search for relevant web data from what we already know in our business register

Deliverable 3.11:

UC5: Report on methodology and results to use online data for business register enhancement

Selective / Statistical scraping: high level view



https://github.com/SNStatComp/SSIG

Def 1.1: *Statistical scraping is the use of online data starting from a-priori information in the respective statistical domain keeping a clear relation with the statistical context.*

Web Intelligence Network Hackathon

WIN, the hackathon

A call to the Web Data community to help us improve official statistics.

Only 14 days left to enter the WIN Hackathon
Don’t miss out.

Web Intelligence Network

Funded by the European Union

WIN. the hackathon

- An ***online*** challenge of 6 weeks (autumn 2024)
- A call to data scientists to ***help*** interpret web data
- A ***selective scraping*** approach
- Dataset of 4000 urls across 4 countries (PL, NL, DE, AT)
- Challenge: to detect social media presence and ecommerce activity
- Q&A sessions during challence
- Solutions are open source
- **10 teams registered** ☺

Web Intelligence
Network

WIN. the hackathon: setup

- An example of selective scraping

Regional map queries:
- NL, PL, DE, AT
- Selective in
  - regions
  - type of activity

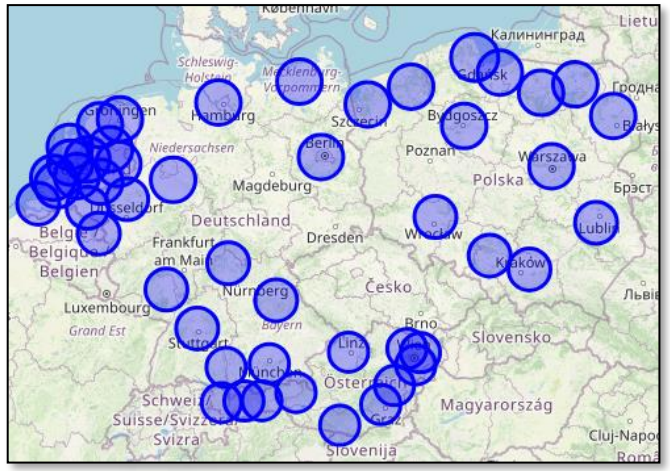
→ ~30 000 URLs → Check & Deduplicate → ~10 000 URLs → Sample → 4000 URLs
1000 per country

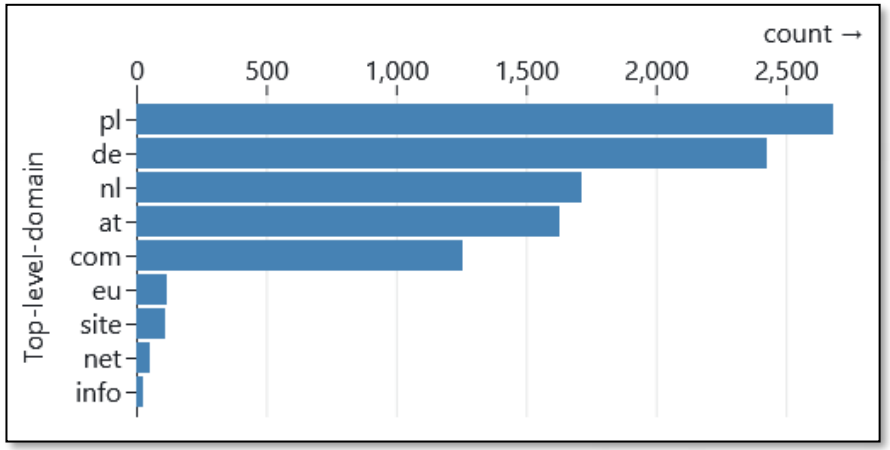
Hackathon challenge:
- 4000 URLS
- Ecommerce
- Social media use:

  fb, linkedin, X, insta, tiktok, YT

manually labeled set
100 per country

Compare





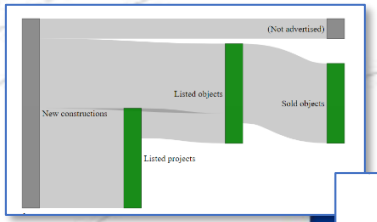
Winners

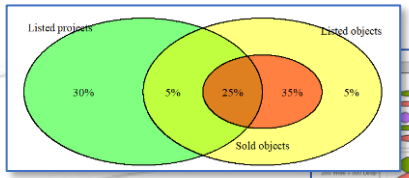
Different approaches, both using AI modeling:

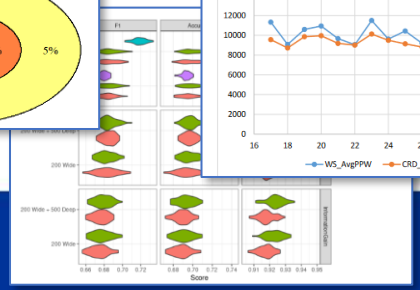
- **Roshna Omer (UNHCR)**
  Enhanced Social Media and E-commerce Detector aka:
  github.com/RoshnaOmer/win-hackathon/

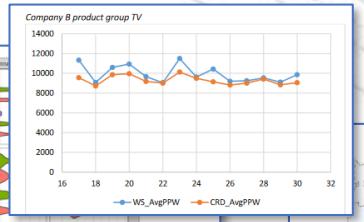
- **Riccardo Corradini, Rita Lima (ISTAT)**
  Freesoftwdreamer team:
  github.com/freesoftwdreamer/Web-Intelligence

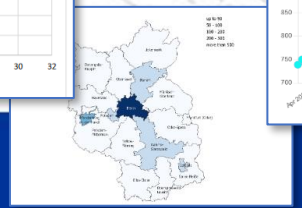
Congratulations and happy to hear their experiences !!!

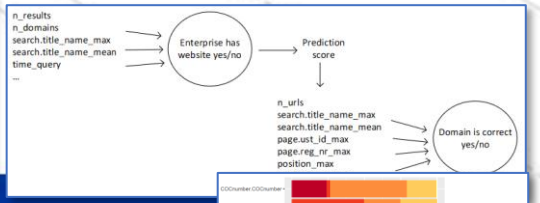


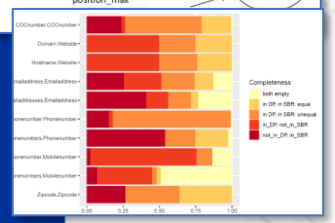


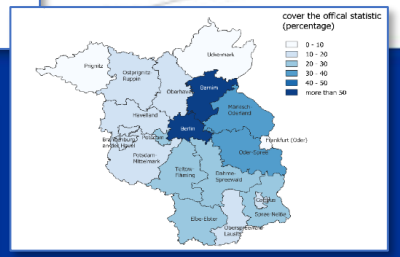


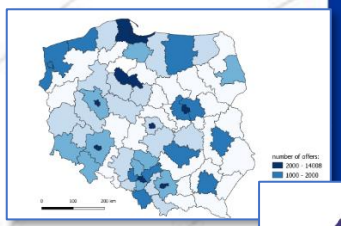


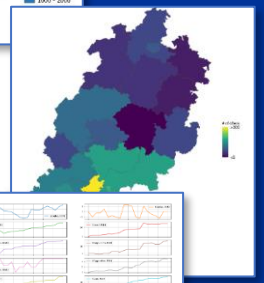












Year	Month	Objects in Poland	Avg. Price in Poland	Objects in Bulgaria	Avg. Price in Bulgaria
2023	January	8276	259,11	3702	252,44
2023	February	9233	259,39	4888	272,00
2023	March	8993	302,74	4812	341,86
2023	April	10017	300,15	4397	305,61
2023	May	10687	314,45	4579	330,20
2023	June	10721	310,02	3506	310,88
2023	July	6886	411,97	1933	394,55
2023	August	8499	376,28	2645	376,06
2023	September	9490	365,03	2910	321,77
2023	October	11573	311,24	3454	320,90
2023	December	10396	300,44	3045	308,71

Thank you

Thanks to many project partners

Olav ten Bosch: o.tenbosch@cbs.nl

https://github.com/WebIntelligenceNetwork/Deliverables

Web Intelligence Network

Funded by the European Union