WORLD STATISTICS CONGRESS 2025 THE HAGUE

Centraal Bureau voor de Statistiek

International Statistical Institute

Official statistics using web data: new use cases

Olav ten Bosch
Statistics Netherlands

Wednesday 8 Oct., 10:50AM - 12:30AM

Contents

- Web data in official statistics
- WIN-WP3: new use cases
- A hackathon
- Statistical Scraping

WORLD STATISTICS CONGRESS 2025 THE HAGUE

International Statistical Institute

WEB DATA IN OFFICIAL STATISTICS

Why web data in official statistics?

Administrative sources

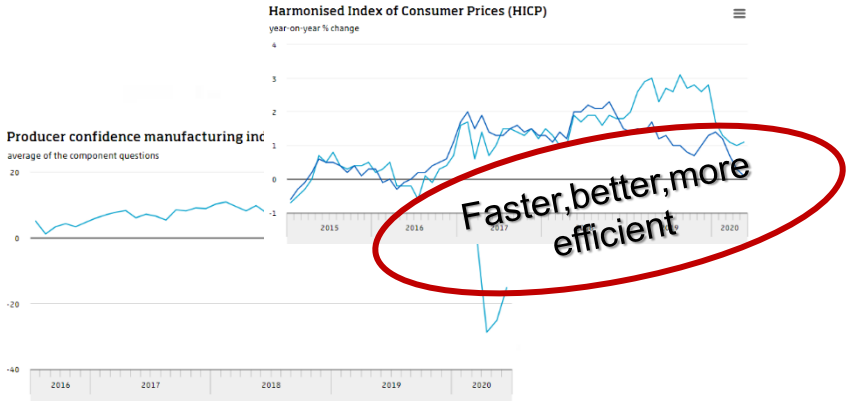
- Tax, social security
- Municipalities/ Provinces
- Supermarkets
- …

Enhancing registers

Statistical population discovery

Internet sources

- …
- Surveys Less!!!



New indicators

Expected

Order position among exporters
today 15:00

Bankruptcies during the coronavirus crisis
tomorrow 12:00

Monitor of Well-being & SDGs
03/07/2020 00:00

Natural capital in the Netherlands (SDG)
03/07/2020 14:00

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 (Textkernel), 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

Big variety in UCs

WORLD STATISTICS CONGRESS 2025 THE HAGUE

International Statistical Institute

WEB INTELLIGENCE NETWORK (WIN)
WP3: NEW USE CASES

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

Deliverables: https://github.com/WebIntelligenceNetwork/Deliverables/tree/main/WP3

Challenges


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

UC1: characteristics of the real estate market

Aim:

- to monitor the ***real estate market*** that responds ***quickly*** to the economic cycles and is ***not fully covered by administrative data***

Potential use:

- Flash estimates
- In-depth research, hedonic indices, modeling real estate characteristics

Landscaping phase:

- Each candidate portal was assessed using a standardised ***checklist*** for assessment of data sources

- Consistent indicators across countries :

Common basic indicators:

- number of offers
- average price per square meter (sale)
- share of offers by price per square meter classes (sale)
- average surface area in square meter
- share of offers by surface area classes
- average number of rooms
- share of offers by number of rooms
- average price (rent)
- share of offers by price classes (rent)

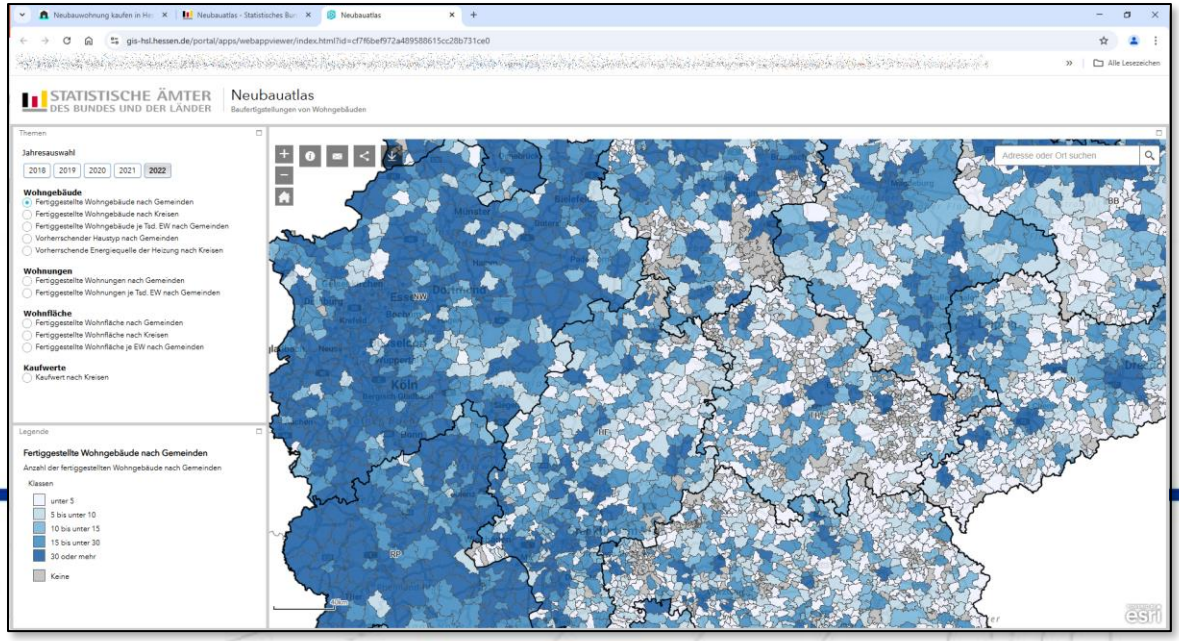
Real-estate statistics from web data
Klaudia Peszat

Next presentation !

UC2: Measuring construction activities using advertisements from real estate portals

Goal:

- From ads appearing at online real estate portals: can **we** produce some **meaningful early estimate** of the number of **newly constructed buildings** that have **become available** in a **specific year** of reference?



Web Intelligence Network

UC2: Measuring construction activities using advertisements from real estate portals (2)

Summary

- Successfully collected data over a longer period of time (up to three years weekly / monthly data)
- … meaningful number… ? Yes!
- Comparable to official statistics?
  - Only aggregates, no linkage of microdata
  - Different concepts / „specification error“
  - Undercoverage, missing data
  - Duplicates
  - Bias (higher coverage of ads in urban areas than in rural areas)

Tobias Gramlich: https://win2025.stat.gov.pl/Presentations

UC3: online prices of household appliances and audio-visual, photographic and information processing equipment

Results:

- ***Weekly*** price collection in ***Bulgaria*** (4 data sources initially, later 3) and ***Sweden*** (2 online sources) to calculate average prices and ***price indices*** over ***longer period*** (>2 yrs) for specific product categories (blenders, steam irons, coffee machines, TVs, washing machines).
- Swedish data integration pilot on combining ***web data*** with ***cash register*** data to estimate ***weights*** of products, products groups or COICOPs, where this information is not available.
- Proof of concept executed on laptops and TVs and cash register data from 2 companies.
- Open issues: weight estimation algorithm, bias in web data, scaling up, scaling up of method, timing of CPI data versus cash register dataflow,
- Further input welcome!

Deliverable 3.7:
UC3: Report on methodology and results for online prices

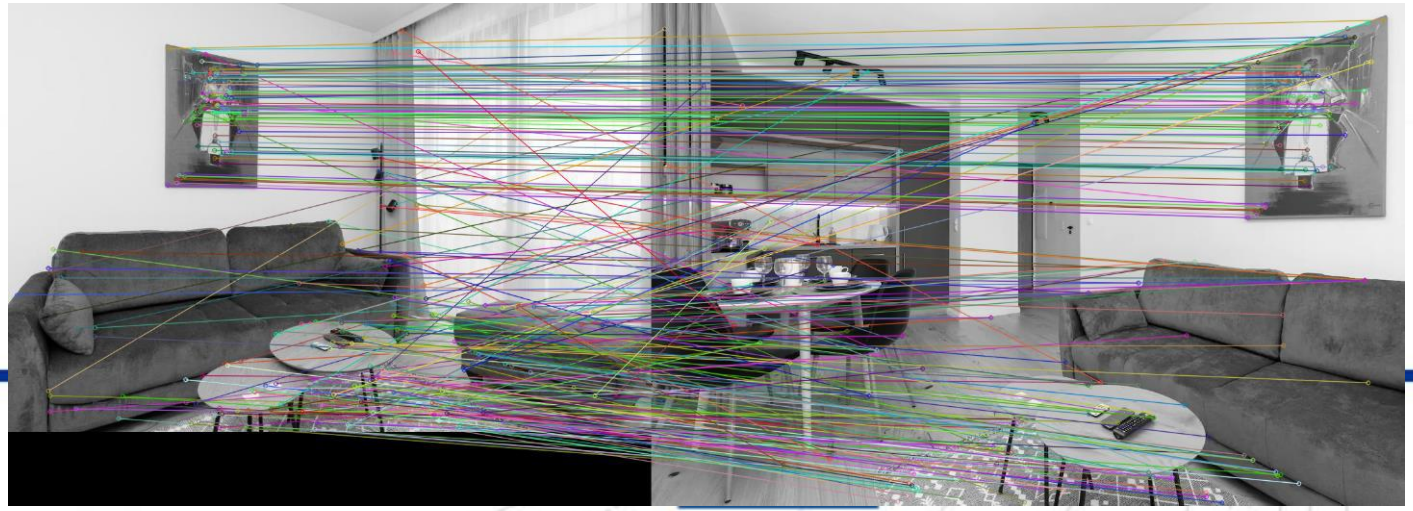
UC4: experimental statistics in tourism (1)

Aim:

- to develop experimental indicators based on data collected through web scraping from online platforms for the purpose of conducting statistical research in the field of tourism
- Use of web data for ***accommodation base*** in tourism (supply side of tourism) and ***tourists' travel patterns and expenditures*** (demand side of tourism)
- Web data can be used for ***validating and imputing*** missing records in sample surveys of tourist travel and spending (demand side of tourism)

Deduplication methodology:

- SIFT (Scale-Invariant Feature Transform) algorithm

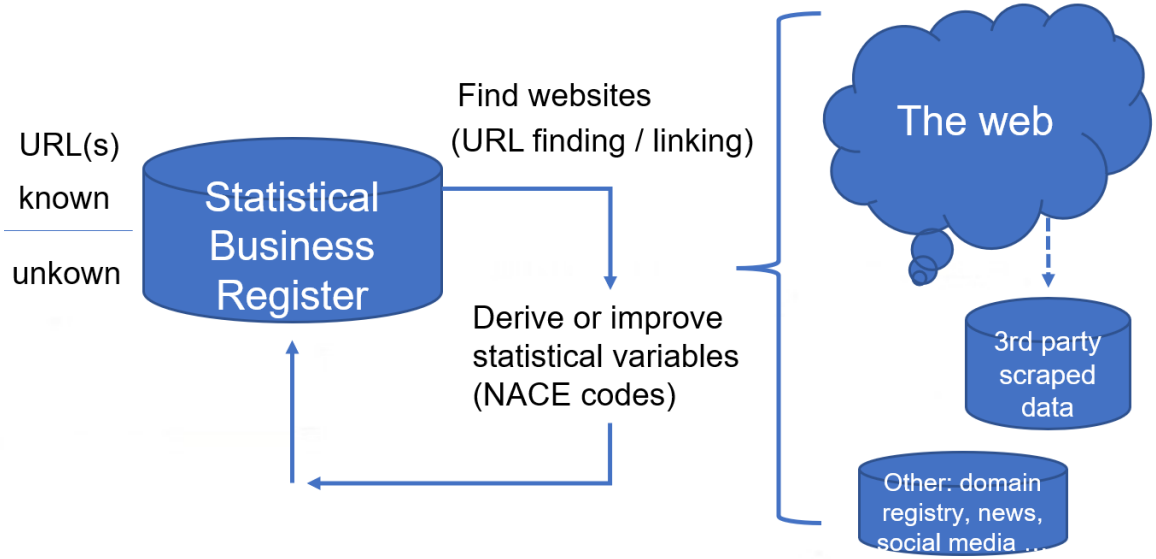


Web Intelligence Network

UC4: experimental statistics in tourism (2)

- Using web scraping repeatedly, it is possible to obtain data on accommodation ***rental prices*** for specific tourist ***destinations*** and during a specified ***period of time***.
- Indicators regarding the ***average rental prices of accommodations*** in ***Poland*** and ***Bulgaria*** have been developed. The rental prices exhibited distinct ***seasonality***.
- This data is also useful for ***validating*** and ***imputing missing data*** in ***sample surveys*** on travel and tourist expenditures conducted by NSIs within the EU.
- Comparing images across platforms that offer accommodation bookings can assist in the ***deduplication*** of accommodation property datasets.
- The research demonstrates potential in utilizing web scraping techniques to estimate ***travel-related expenditures***. While the methodology is still under refinement, the initial findings are promising

UC5: business register enhancement



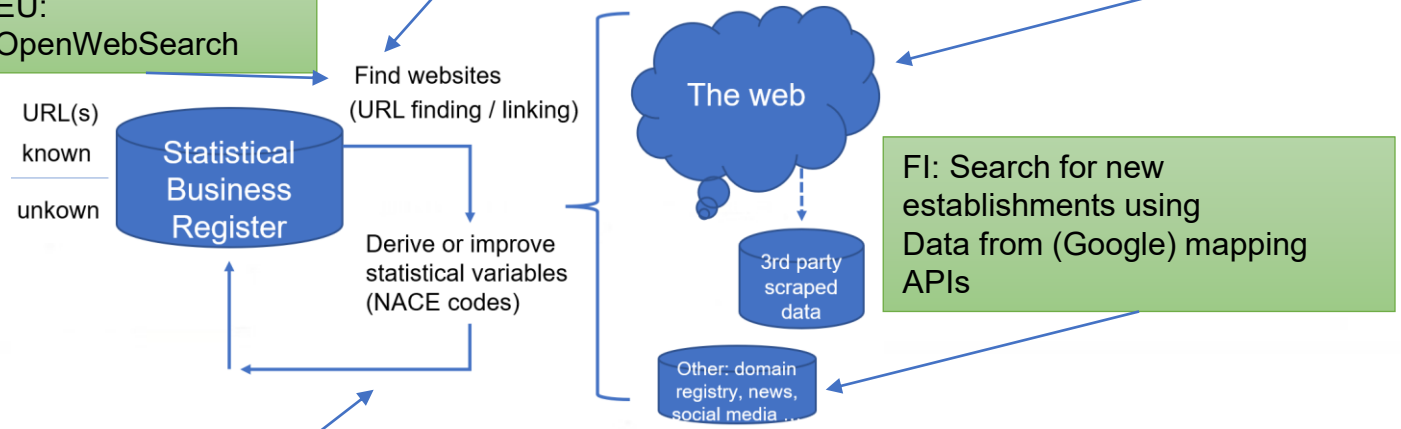
We search for relevant web data from what we already know in our business register

UC5: business register enhancement

HESSE/AT/NL/SE: URLfinding via search engines (Google API)
FI: URL finding via domain registry API

AT, NL: NACE detection

EU:
OpenWebSearch



FI: Search for new establishments using
Data from (Google) mapping APIs

Contact information discovery (Hesse)
- extracting email addresses
- names of business executives

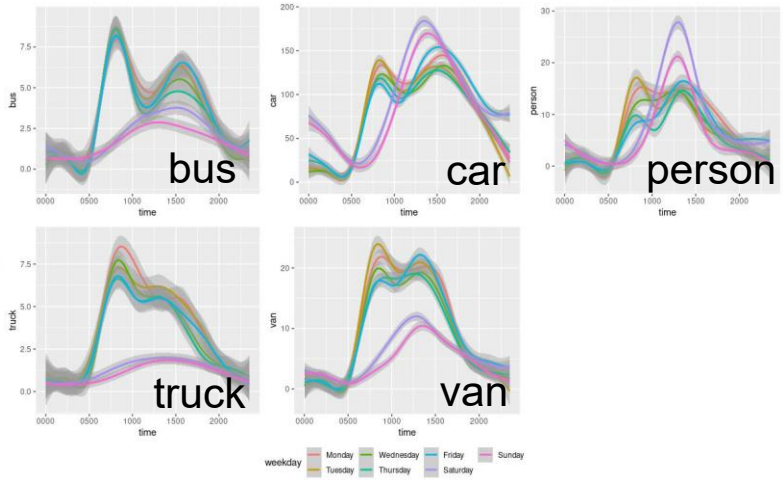
Deliverable 3.11:

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

UC6: faster economic indicators, traffic cams

Aim:

- to explore the applicability of using ***publicly available traffic camera images*** to calculate a ***busyness indicator***, successfully piloted in the UK, and to adapt it for use in other countries, with ***Sweden*** as a case study



- Method is technically & methodologically feasible and portable across countries
- Different weather conditions

However:

- Evolving social attitudes towards privacy => less camera's images available

Future: combine with:

- Combine such data with MNO / sensor data / citizen science initiatives (i.e. telraam)

WORLD STATISTICS CONGRESS 2025 THE HAGUE

International Statistical Institute

WEB INTELLIGENCE NETWORK (WIN)
A HACKATHON

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 ***statistical scraping*** approach
- Dataset of 4000 enterprise 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** ☺

WIN. the hackathon: setup

- An example of statistical scraping

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

→ ~30 000 URLs

Check & Deduplicate

→ ~10 000 URLs

Sample

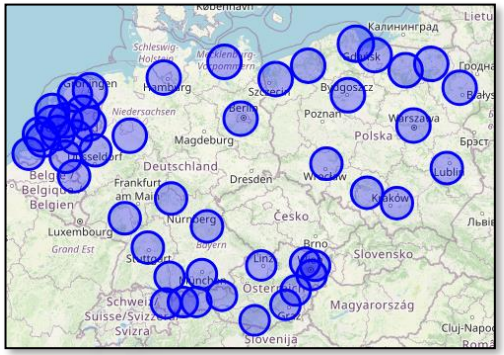
→ 4 000 URLs
1000 per country

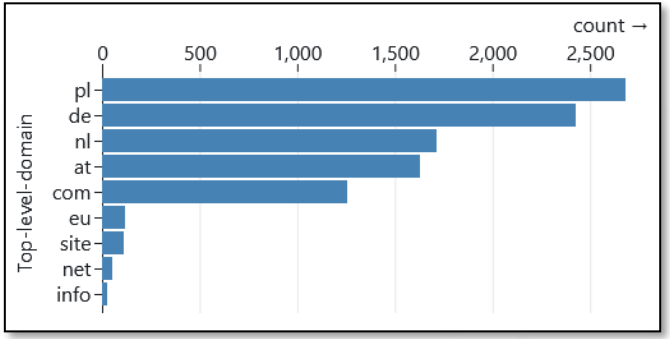
Hackathon challenge:
- 4000 URLS
- Ecommerce
- Social media use:

  fb, linkedin, X, insta, tiktok, YT

manually labeled set
100 per country

Compare





Winners awarded at NTTS 2025
Both used AI

WORLD STATISTICS CONGRESS 2025 THE HAGUE

International Statistical Institute

STATISTICAL SCRAPING

A CONCEPT

Web data: typical discussion

Hey Olav, can we get the data from sites X, Y and Z?

Technically yes, but do we need it all?

Yes, just give me everything, then I'll analyse, link and aggregate

But the web - and this website – both offer excellent query possibilities, you could take a subset of meaningful data

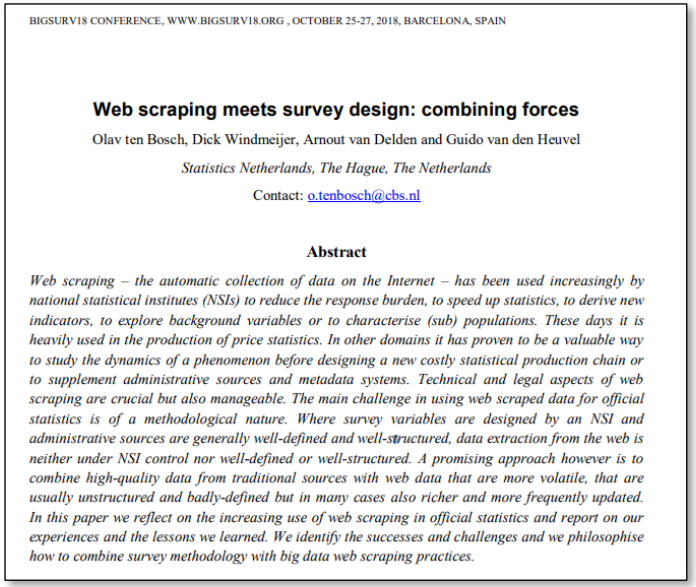
Yeah, but why should we do that?

*If you can **query** for webdata matching a statistical unit (or aggregate) then you can do 'statistical scraping', and use traditional quality indicators*

Oh great. Hmm, but what if I don't know the population?

In that case you might need explorative scraping in addition. Please think about it?

Back in 2018



BIGSURV18 CONFERENCE, WWW.BIGSURV18.ORG , OCTOBER 25-27, 2018, BARCELONA, SPAIN

Web scraping meets survey design: combining forces

Olav ten Bosch, Dick Windmeijer, Arnout van Delden and Guido van den Heuvel

Statistics Netherlands, The Hague, The Netherlands

Contact: o.tenbosch@cbs.nl

Abstract

Web scraping – the automatic collection of data on the Internet – has been used increasingly by national statistical institutes (NSIs) to reduce the response burden, to speed up statistics, to derive new indicators, to explore background variables or to characterise (sub) populations. These days it is heavily used in the production of price statistics. In other domains it has proven to be a valuable way to study the dynamics of a phenomenon before designing a new costly statistical production chain or to supplement administrative sources and metadata systems. Technical and legal aspects of web scraping are crucial but also manageable. The main challenge in using web scraped data for official statistics is of a methodological nature. Where survey variables are designed by an NSI and administrative sources are generally well-defined and well-structured, data extraction from the web is neither under NSI control nor well-defined or well-structured. A promising approach however is to combine high-quality data from traditional sources with web data that are more volatile, that are usually unstructured and badly-defined but in many cases also richer and more frequently updated. In this paper we reflect on the increasing use of web scraping in official statistics and report on our experiences and the lessons we learned. We identify the successes and challenges and we philosophise how to combine survey methodology with big data web scraping practices.

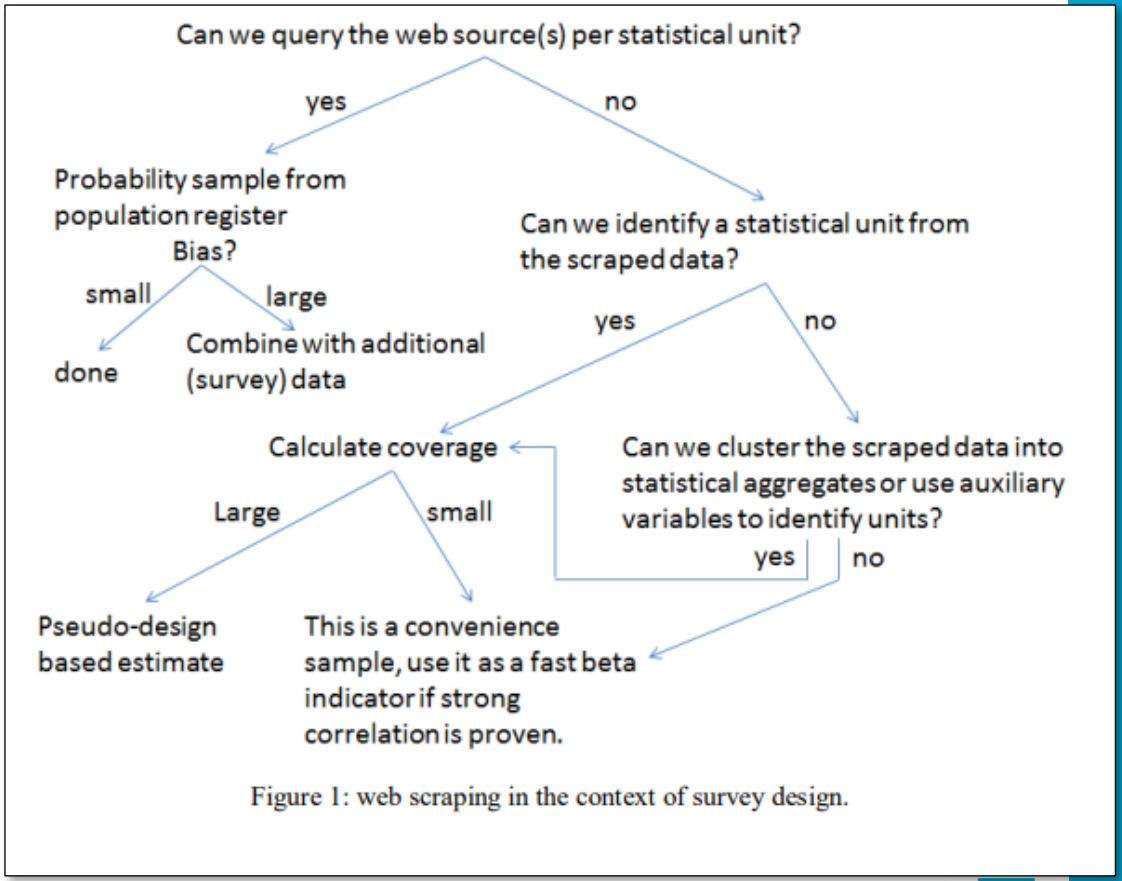


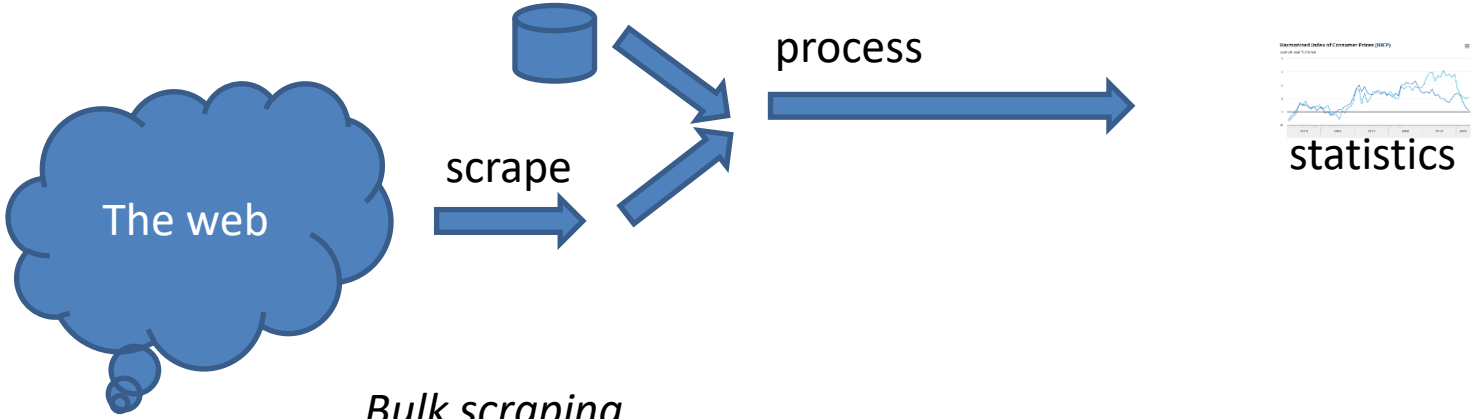
Figure 1: web scraping in the context of survey design.

General workflow for web data, 1st try (2018)

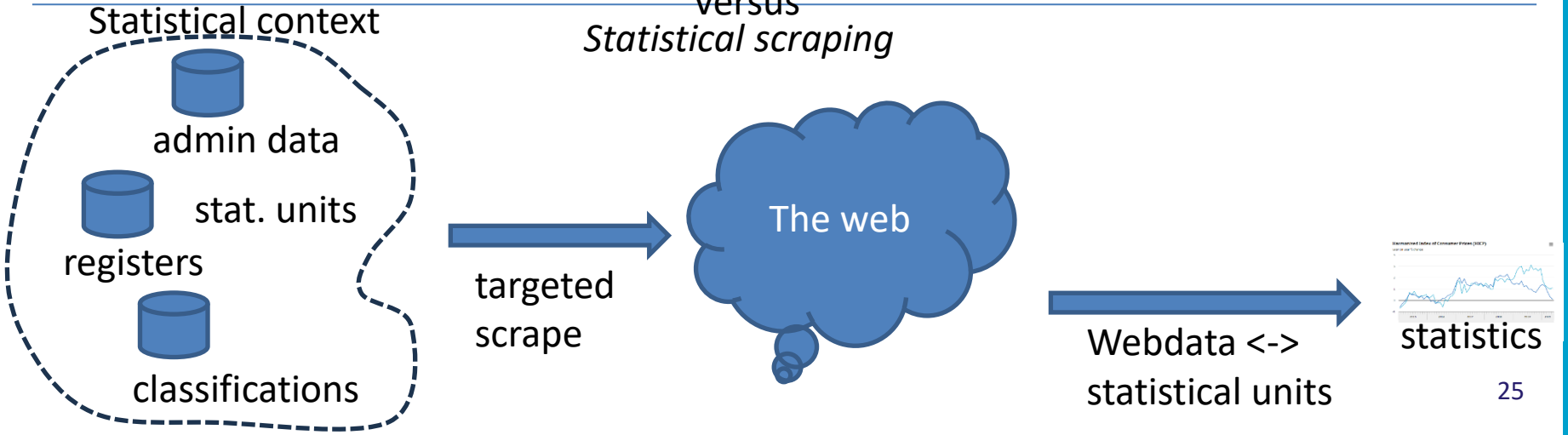
BigSurv2018

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

admin data, registers, stat. units, classifications



Bulk scraping
versus
Statistical scraping



Definition (Q 2024)

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.*

- Methodological consequences:
- In general, statistical scraping ***helps*** cope with different types of ***representation errors***
- If applied on unit level it becomes possible to calculated proven survey methodology ***quality indicators***

- Other aspects:
- A targeted scrape leads to ***smaller***, ***more manageable*** data streams
- Web queries may need ***possibly sensitive*** statistical ***input data***, which should be handled with care

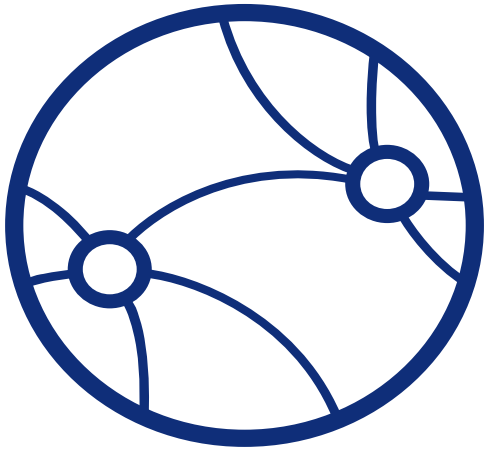
Statistical Scraping Interest Group (SSIG) Meetings

SSIG1: 16-17 sep 2025 in Vienna

SSIG2: 15-16 april 2026 in The Hague

SSIG3: Autumn 2026 in ?

SSIG1: Spring 2027 in ?



Statistical Scraping
Interest Group

https://github.com/SNStatComp/SSIG

WORLD STATISTICS CONGRESS 2025 THE HAGUE

Centraal Bureau voor de Statistiek

International Statistical Institute

THANK YOU.

QUESTIONS, IDEAS, SUGGESTIONS WELCOME

O.TENBOSCH@CBS.NL
OLAVTENBOSCH.GITHUB.IO