Statistical Scraping: history, concept, where we are, SSIG

Olav ten Bosch with input from many
Statistics Netherlands

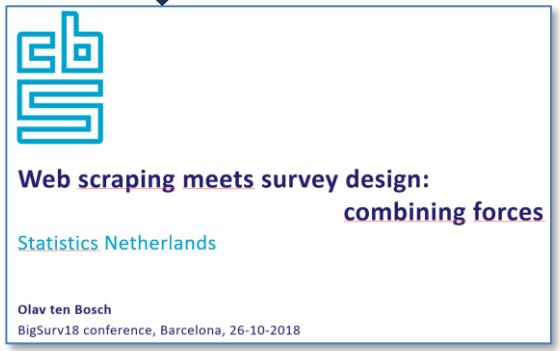
Statistical Scraping Interest Group meeting1 (SSIG1), 16-17 sep, Vienna
https://github.com/SNStatComp/SSIG

Statistical Scraping
Interest Group

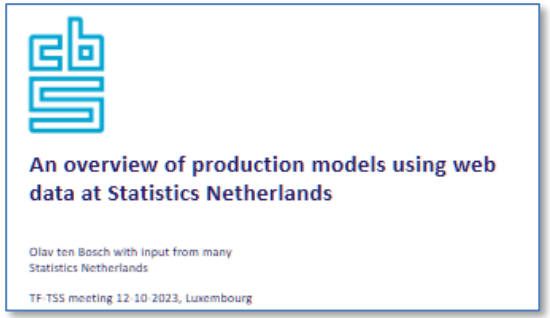
History

History

?



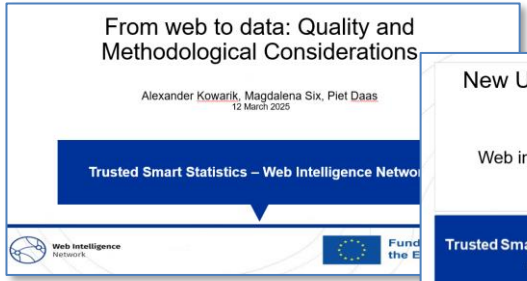
BigSurv 18

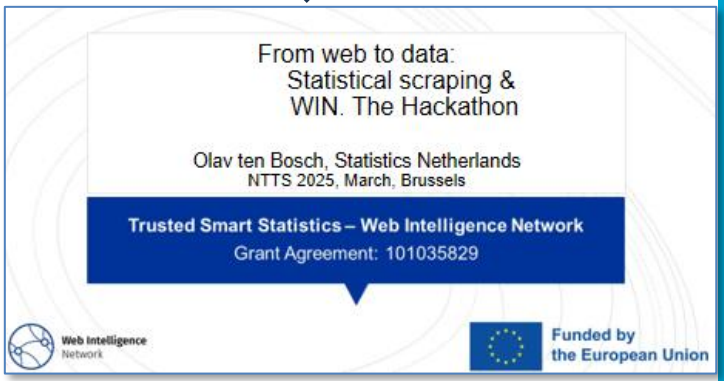


Eurostat TF-TSS 23

Qconf 24

ESSnet WIN
Gdansk 25





NTTS 25

WEB-FOSS-AT / WEB-FOSS-NL

SSIG

>15 years of web data at CBS

2008-2010
Fuel prices
Real estate
Airtickets

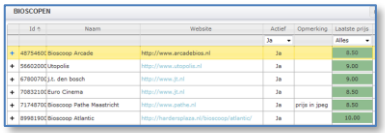
2011-2019
Experimenting
towards offstats

- ***Webshops***: CPI (inflation): prices (clothing), books, travel, consumer electronics
- ***Enterprise websites***: e
  social media use, NAC
  businesses, drone com
  platform economy
- ***Annual reports***: financ
- ***Social media***: social te
  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

Manual retail price observations discontinued

13/01/2020 14:00

2020



Semi-automatic scraping
2012 -> ongoing

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.

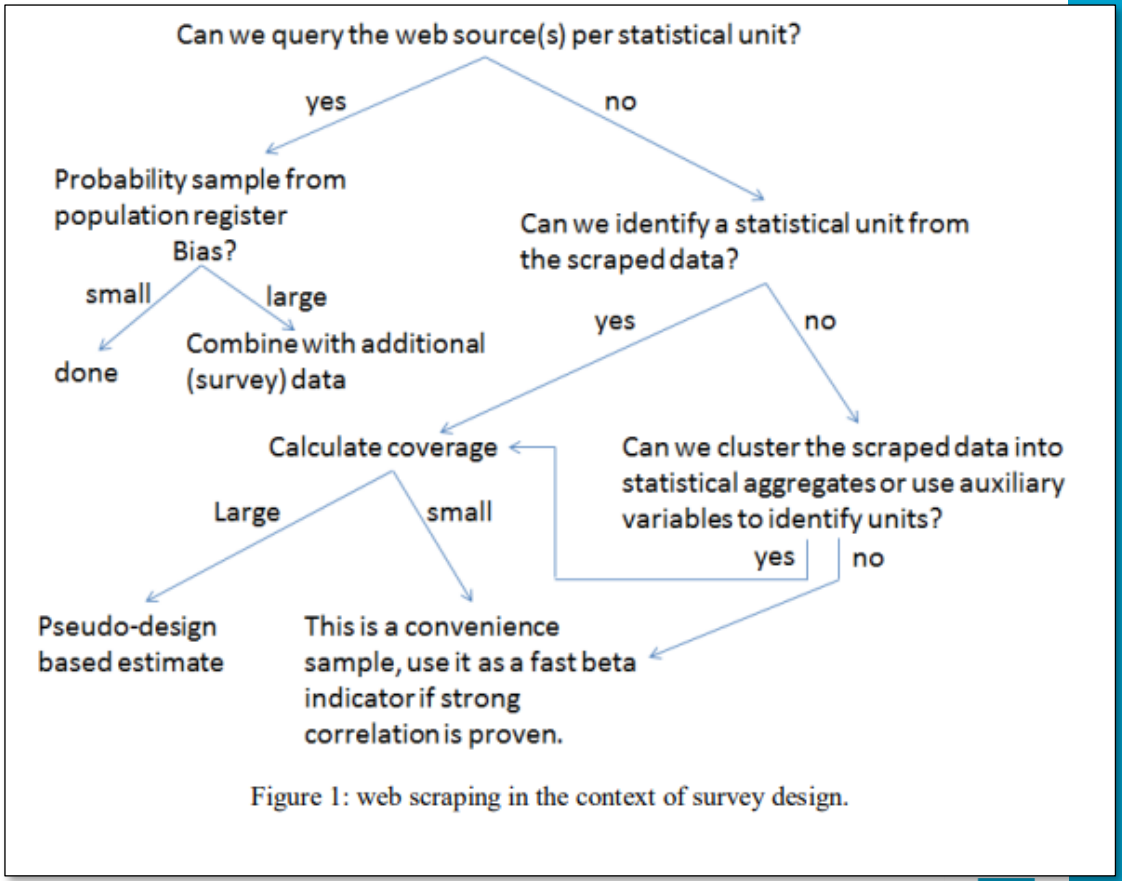


Figure 1: web scraping in the context of survey design.

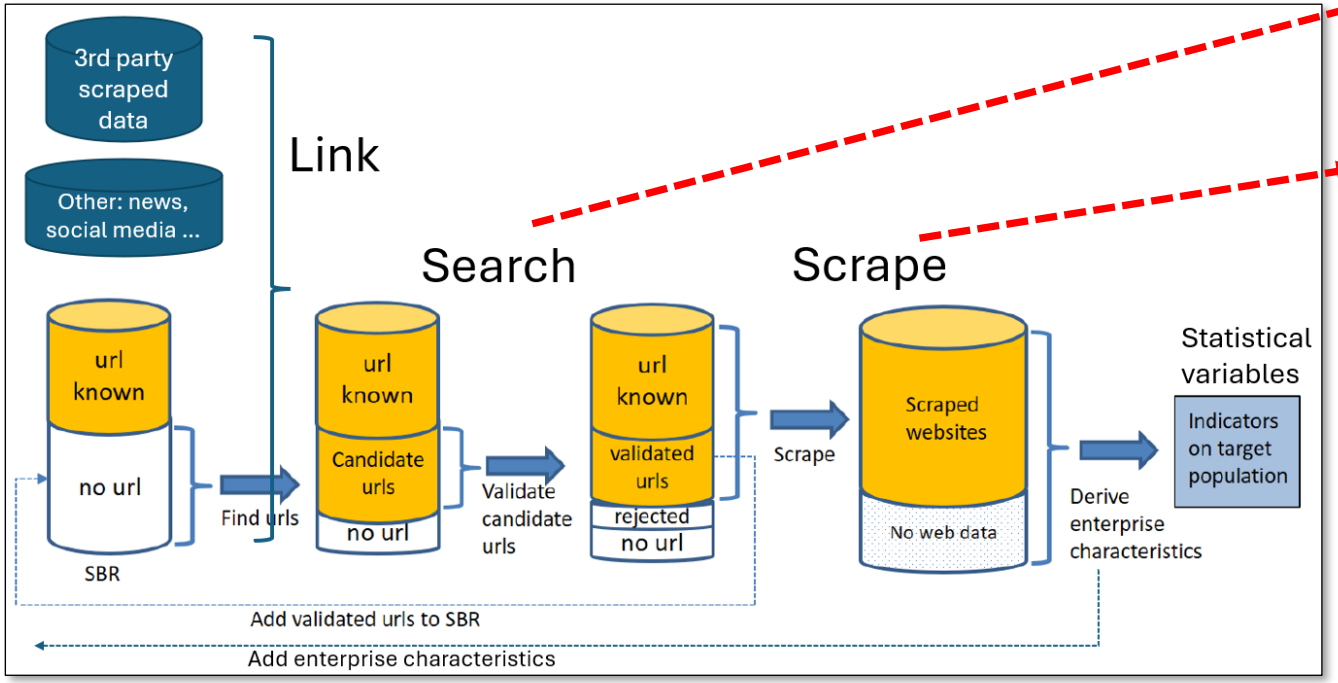
General workflow for web data, 1st try (2018)

BigSurv2018

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

The conference is partly financed by the European Union

Example: Business register enhancements



using ***search engine***(s): query on enterprise name / details

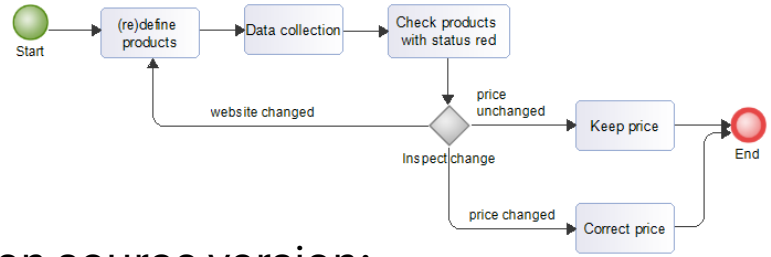
focused scraping: focus on interesting parts of web pages, like 'about us page' etc.

Both are target scrapes: It starts from what we already know in the business register

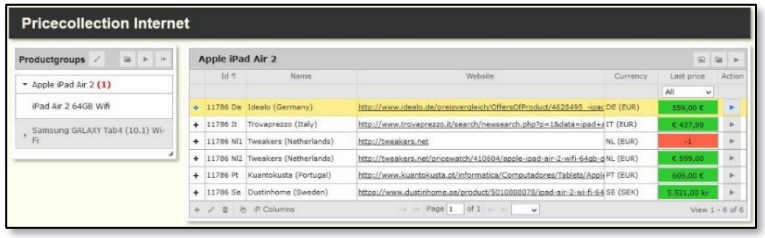
The conference is partly financed by the European Union

Other examples: Robot-assisted price collection

- ***Robottool***: (2012->ongoing) 8 users; 2850 price observations/month
- ***Check*** products with infrequent price changes ***easily***:
  - Examples: Cinema tickets, drivers lessons, car / bike repair, music instruments, farmacy, snackbars, dentists, sports, museum
- Price specialists define ***path*** to price and product to be checked



Green: nothing changed -> last price saved
Red: needs attention



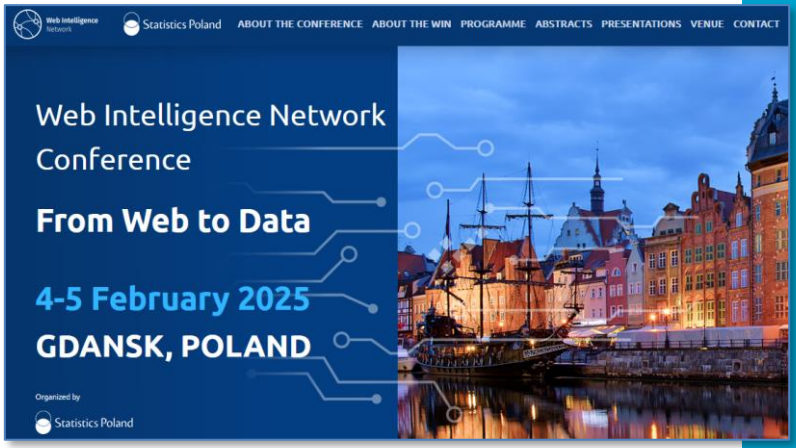
Open source version:
https://github.com/SNStatComp/RobotTool

The conference is partly financed by the European Union

Other examples

- ***Tourism***: for each hotel star rating category a number of hotels are selected from the national ***star rating overview site*** (sample design). This sample is then scraped regularly at ***low frequency***.
- ***Education***: visit school web sites starting from a register of schools to draw a sample of teachers. Collecting data on a ***representative subset*** of doctorate holders.
- ***Internet standards***: from ICT survey automaticaly derive state of play if use of internet standards at enterprise websites.
- ***Platform economy***: from ICT survey + webscraping identify tate of play of platform economy sites (Airbnb, Uber, eBay)
- ***Labour market***: is statistical scraping applicable? Start from business register and visit enterprise websites and possibly job portals, keeping the relationship with the statistical unit?

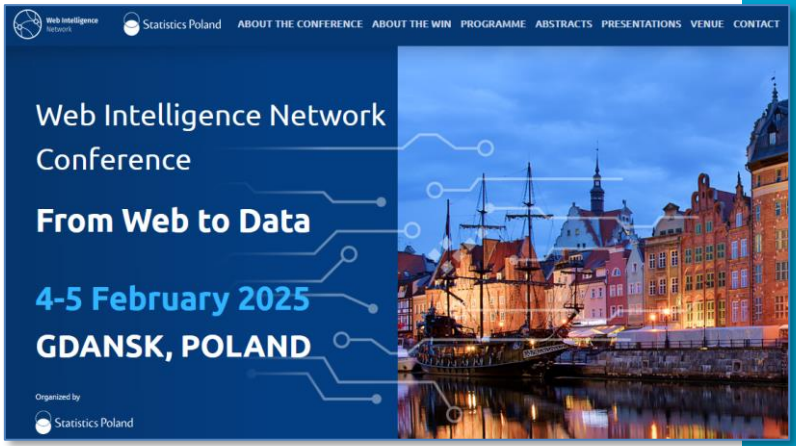
WIN final conf (Gdansk 25)



WIN topics:

- OJAs
- OBEC (ICT survey based)
- New use cases
  - **UC1** Characteristics of the real estate market — **PL**, BG, DE-HSL, DE-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
- Quality & Methodology on using web data

WIN final conf (Gdansk 25)



Some other topics:

- Firms innovation capabilities
- Green skills analysis
- Combining OJAs with probability sample data
- Applying survey sampling theory to web-scraped data
- Estimating interregional trade using weblinks
- Web data for energy statistics
- Estimating vehicle mileage & analyzing road traffic accidents
- The use of AI in web data (Marko Grobelnik)

Conclusion Alex WIN-WP4:

- Still big challenges in using web data for proper statistical conclusions, but we are getting closer
- Proper methodology for design, validation and estimation is needed
  - Quick and dirty is only a solution for experiments

All deliverables at https://github.com/WebIntelligenceNetwork/Deliverables

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** ☺

WIN. the hackathon: setup

- An example of selective scraping

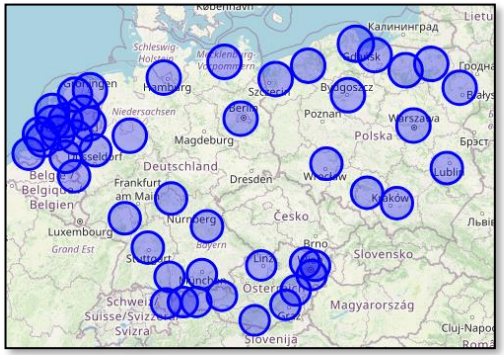
Regional map queries:
- NL, PL, DE, AT
- Selective in
  - regions
  - type of activity

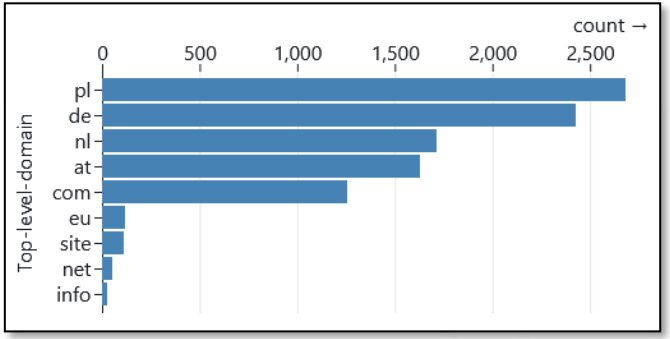
→ ~30 000 URLs → Check & Deduplicate → ~10 000 URLs → Sample → 4000 URLs
1000 per country

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

  fb, linkedin, X, insta, tiktok, YT

manually labeled set
100 per country → Compare





A Statistical Scraping experiment: spatial sample

Concept

Concept

admin data, registers, stat. units, classifications

process

statistics

The web

scrape

CBS-Prices/bulk scraping
ESTAT-OJAs

Big data approach
versus
Statistical scraping

CBS-Job market
CBS-tourism

admin data

stat. units

registers

classifications

targeted
scrape

The web

CBS-Prices/RobotTool
WIN-UC5-SBR enhancements
CBS-platform economy

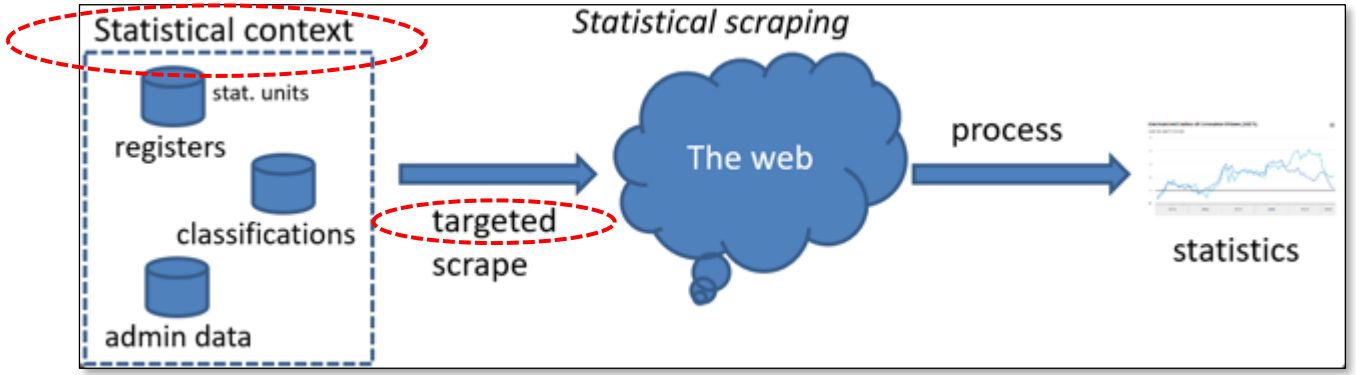
Webdata <->
statistical units

statistics

The conference is partly financed by the European Union

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.*



The conference is partly financed by the European Union

Consequences

Methodological:

- 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:

- 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

Where we are

Statistical Scraping: typical discussion

"Yeah ... ☺"

"But ... 😐"

A more methodological approach to using web data in official statistics

We shouldn't do a survey on big data, we can get it all...

But the web offers query possibilities in many ways, we take exactly what we need...

Not for all web data , for some we can…

*If we **can** find data linked to a statistical unit, then we can do statistical scraping*

Only if we know the population, population discovery still needs bulk scraping

Agree, where is the optimum?

Statistical Scraping in 2025

We need you !

to refine, to invent, add, to experiment

SSIG

Interest group

- Informal
- Learn, share
- Explore
- Examples, examples, examples

- Formalize -> Statistical scraping theory

Thanks

Olav ten Bosch
o.tenbosch@cbs.nl

Thanks to colleagues of all statistical offices that added to the concept of statistical scrping

BTW: don't forget about the awesome list:
awesomeofficialstatistics.org