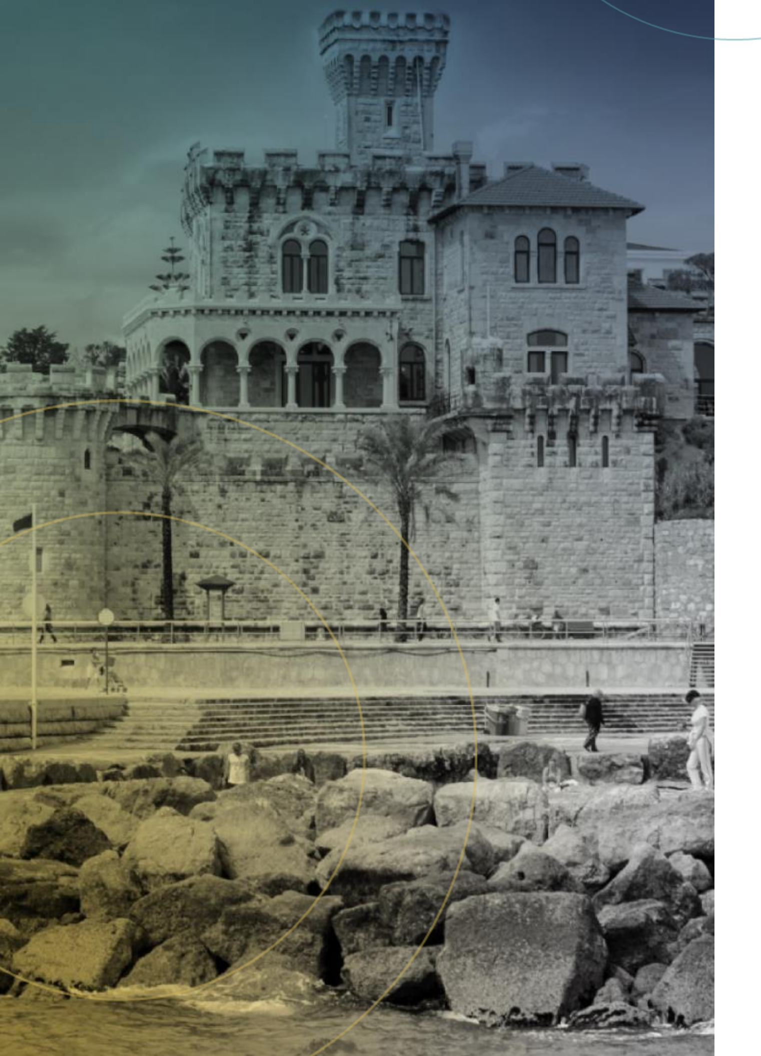
EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL



Statistical scraping: informed plough begets finer crops

Olav ten Bosch, Statistics Netherland

Alexander Kowarik, Statistics Austria

Sónia Quaresma, Statistics Portugal

David Salgado, Statistics Spain

Arnout van Delden, Statistics Netherlands

Instituto Nacional de Estatística
Statistics Portugal

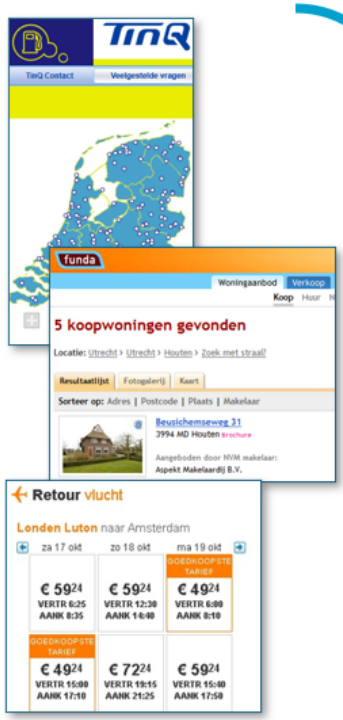
eurostat

The conference is partly
financed by the European Union

Contents

- The use of web data
- Statistical scraping: high level view
- Examples
- Definition, consequences
- Wrap-up

15 years of web data at Statistics Netherlands



TinQ
TinQ Contact | Veelgestelde vragen

funda
Woningaanbod | Verkoop
Koop | Huur

5 koopwoningen gevonden

Locatie: Utrecht > Utrecht > Houten > Zoek met straat?

Resultaatlijst | Fotogalerij | Kaart

Sorteer op: Adres | Postcode | Plaats | Makelaar

Beusichemseweg 31
3994 MD Houten Brochure
Aangeboden door NVM makelaar:
Aspekt Makelaardij B.V.

Retour vlucht

Londen Luton naar Amsterdam

za 17 okt	zo 18 okt	ma 19 okt
€ 59²⁴ VERTR 6:25 AANK 8:35	€ 59²⁴ VERTR 12:30 AANK 14:40	GOEDKOOPSTE TARIEF € 49²⁴ VERTR 6:00 AANK 8:10
GOEDKOOPSTE TARIEF € 49²⁴ VERTR 15:00 AANK 17:10	€ 72²⁴ VERTR 19:15 AANK 21:25	€ 59²⁴ VERTR 15:40 AANK 17:50

2008-2010
Fuel prices
Real estate
Airlines

2011-2019
Experimenting
towards offstats

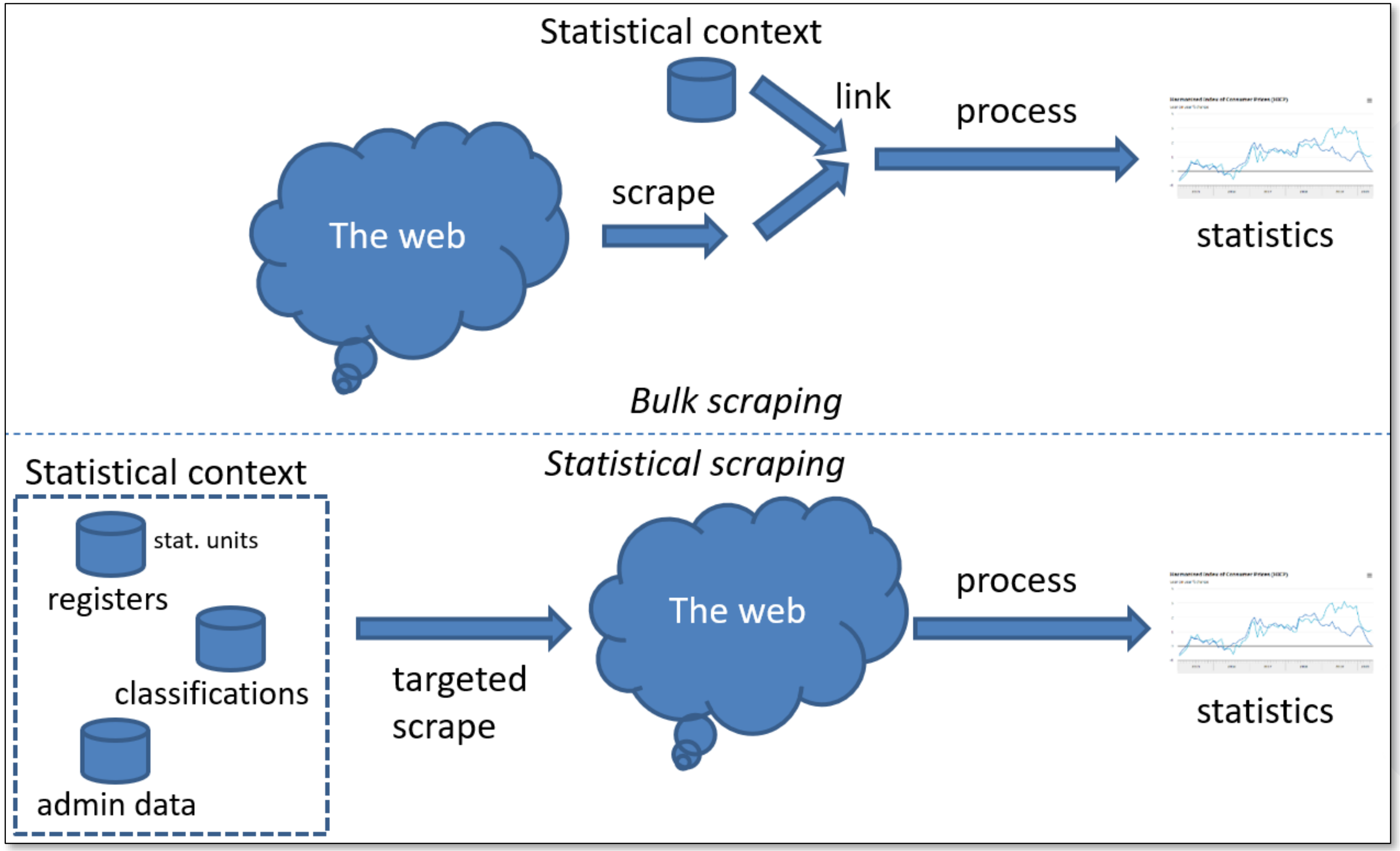
- ***Webshops***: CPI (inflation): prices (clothing), books, travel, consumer electronics
- ***Enterprise websites***: ecommerce activities, webshop detection, social media use, NACE, innovative companies, family businesses, jobs, drone companies
- ***Annual reports***: financial and institutional data
- ***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
- ***Search engines***: enterprises, products

Manual retail price observations discontinued

13/01/2020 14:00

2020

High level view



Statistical context

link

process

scrape

The web

statistics

Bulk scraping

Statistical scraping

Statistical context

stat. units

registers

classifications

admin data

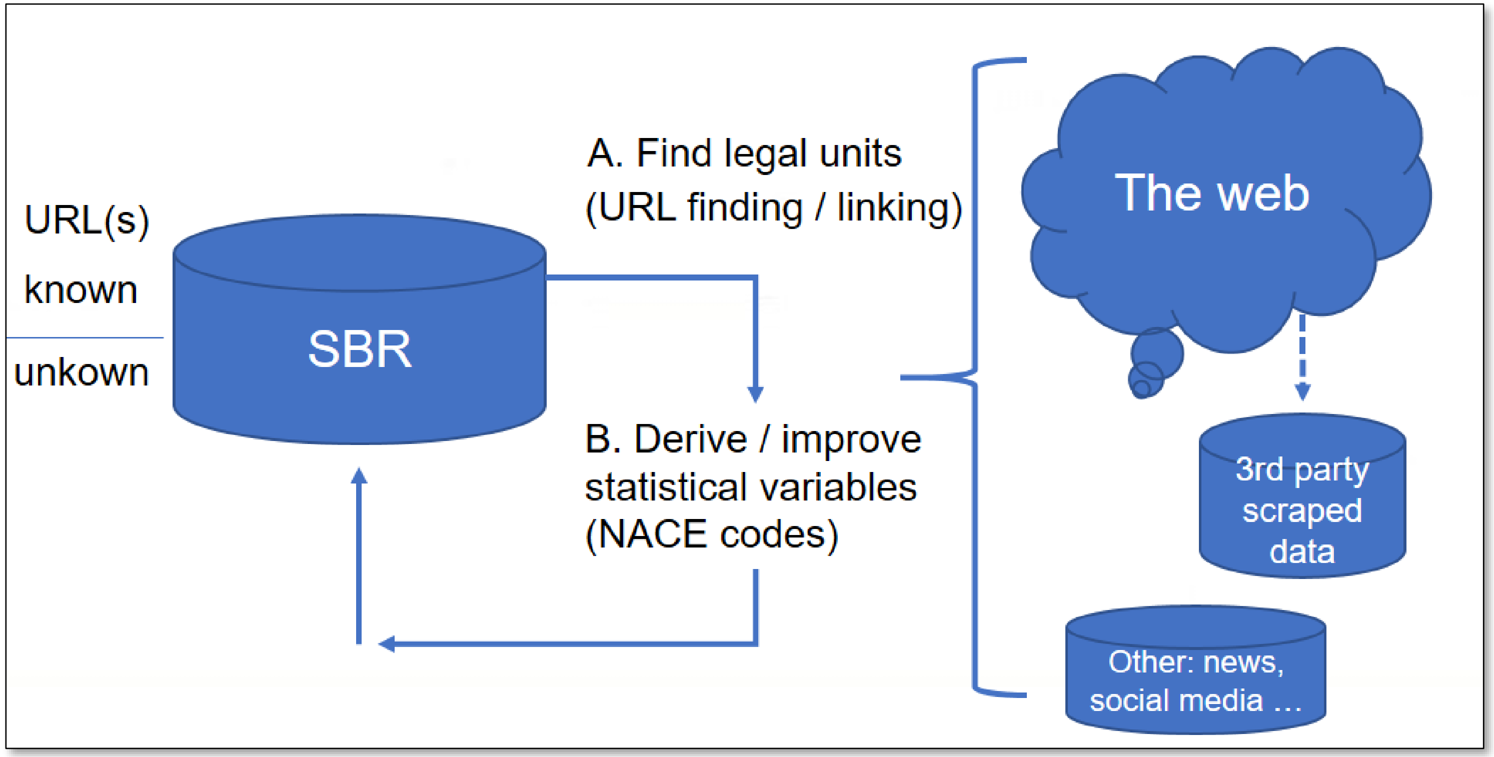
targeted scrape

The web

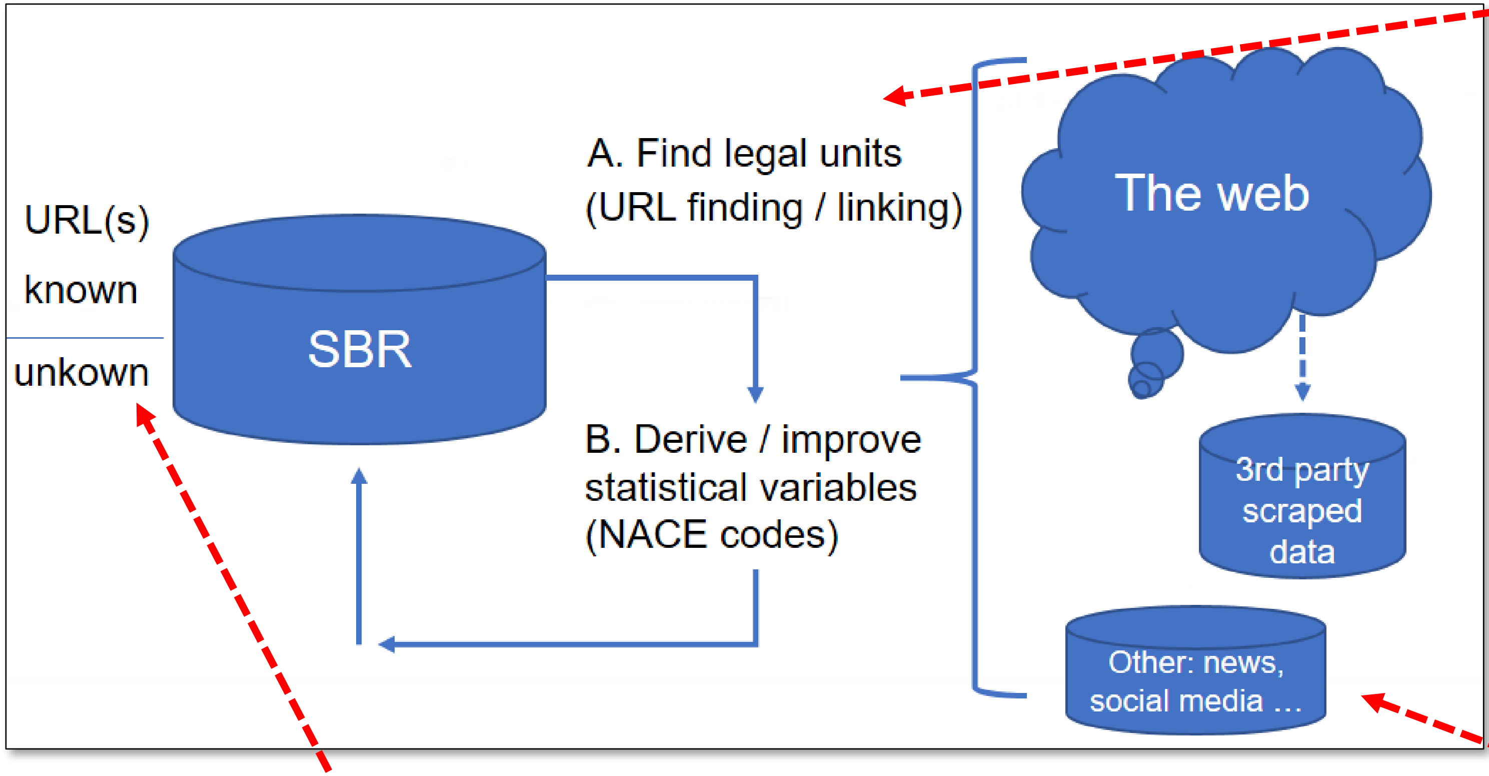
process

statistics

Example: Business register enhancements



Example: Business register enhancements

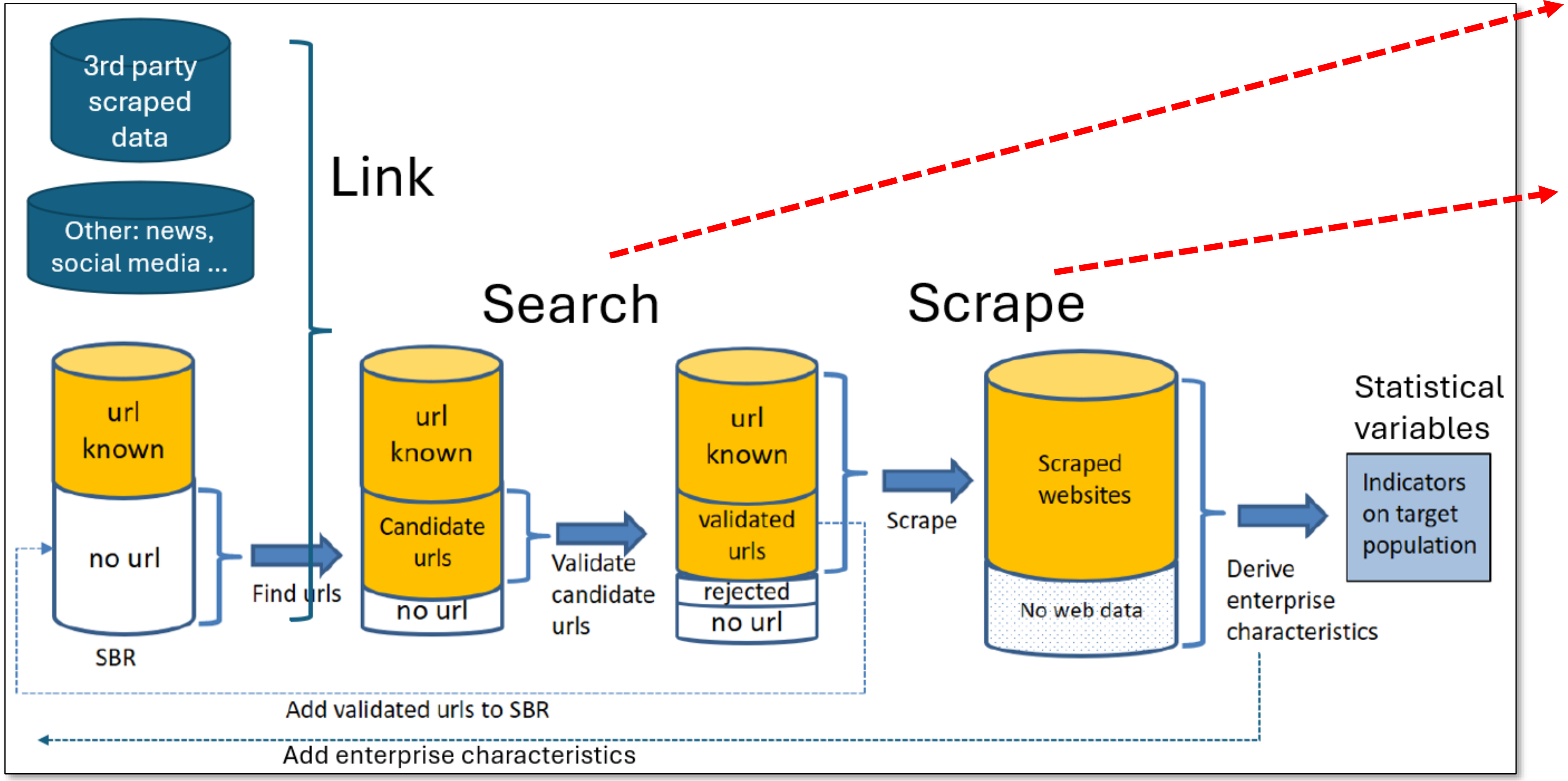


Be aware of n-to-m relationships legal unit and website

The ratio known/unknown is country- specific

Web data is more than websites only

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.

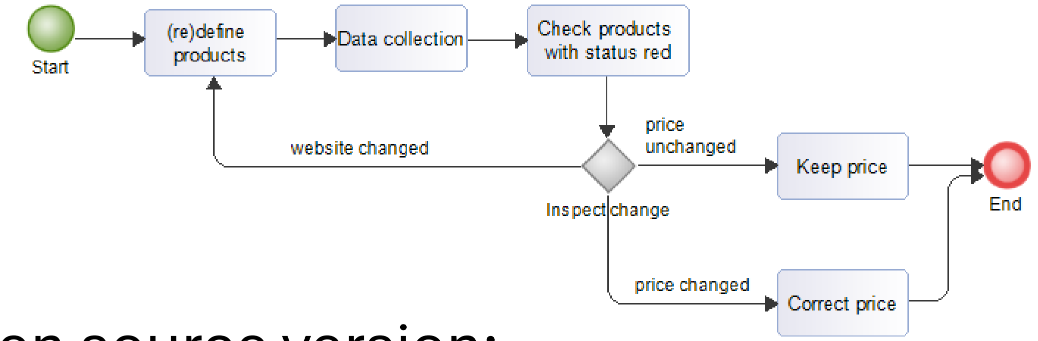
Both are target scrapes: It starts from what we already know in the business register

Example: price statistics

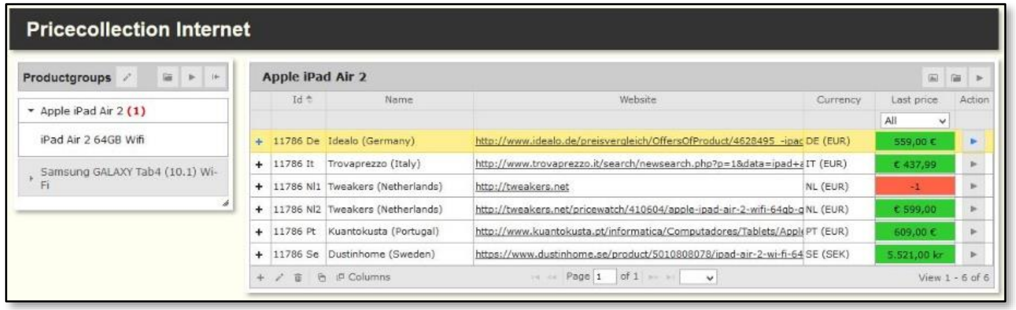
- Early approaches used ***bulk scraping*** (Cavallo, 2009)
- Approaches evolved into a more ***targeted*** scrape of subset of products relating to COICOP categories (statistical context)
- NL: ***Robot-assisted data collection***: basket-based price collection on to the web. For each product a number of measuring spots on the web are defined and regularly visited. In production for over 10 years.
- Even more advanced: use site-specific or general ***search engines*** to discover well-defined products (targeted scrape)

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


Pricecollection Internet

Productgroups

- Apple iPad Air 2 (1)
  - iPad Air 2 64GB Wifi
- Samsung GALAXY Tab4 (10.1) Wi-Fi

Apple iPad Air 2

	Id	Name	Website	Currency	Last price	Action
					All	
+	11786 De	Idealo (Germany)	http://www.idealo.de/preisvergleich/OffersOfProduct/4628495_-ipac	DE (EUR)	559,00 €	▶
+	11786 It	Trovaprezzo (Italy)	http://www.trovaprezzo.it/search/newsearch.php?p=1&data=ipad+a	IT (EUR)	€ 437,99	▶
+	11786 Nl1	Tweakers (Netherlands)	http://tweakers.net	NL (EUR)	-1	▶
+	11786 Nl2	Tweakers (Netherlands)	http://tweakers.net/pricewatch/410604/apple-ipad-air-2-wifi-64gb-g	NL (EUR)	€ 599,00	▶
+	11786 Pt	Kuantokusta (Portugal)	http://www.kuantokusta.pt/informatica/Computadores/Tablets/Apple	PT (EUR)	609,00 €	▶
+	11786 Se	Dustinhome (Sweden)	https://www.dustinhome.se/product/5010808078/ipad-air-2-wi-fi-64	SE (SEK)	5.521,00 kr	▶

Columns Page 1 of 1 View 1 - 6 of 6

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

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.
- ***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?

Generic workflow for webdata

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

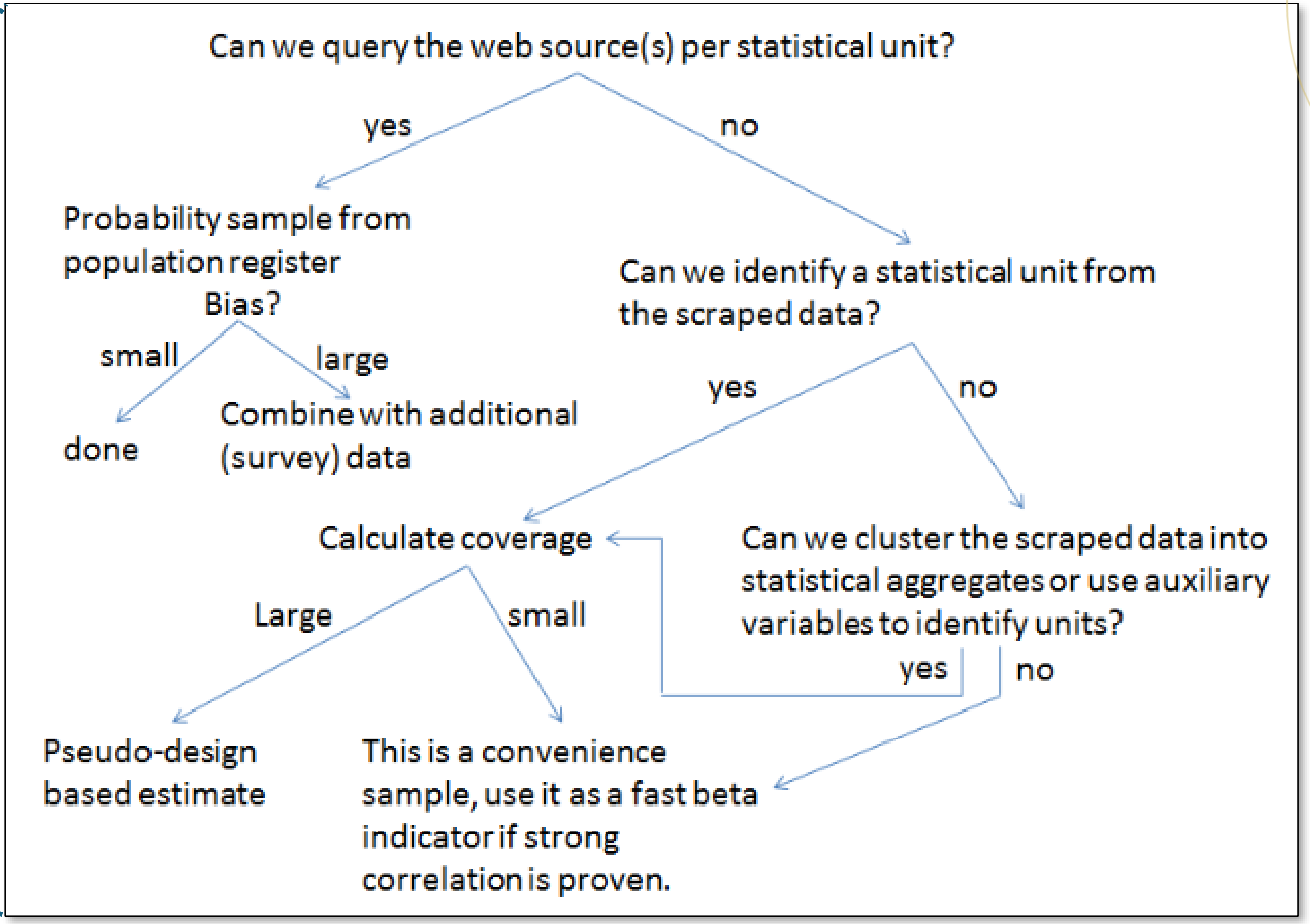
Statistics Netherlands, The Hague, The Netherlands

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.

BigSurv paper, 2018



Can we query the web source(s) per statistical unit?

yes

no

Probability sample from
population register

Bias?

small

large

done

Combine with additional
(survey) data

Can we identify a statistical unit from
the scraped data?

yes

no

Calculate coverage

Can we cluster the scraped data into
statistical aggregates or use auxiliary
variables to identify units?

yes

no

Large

small

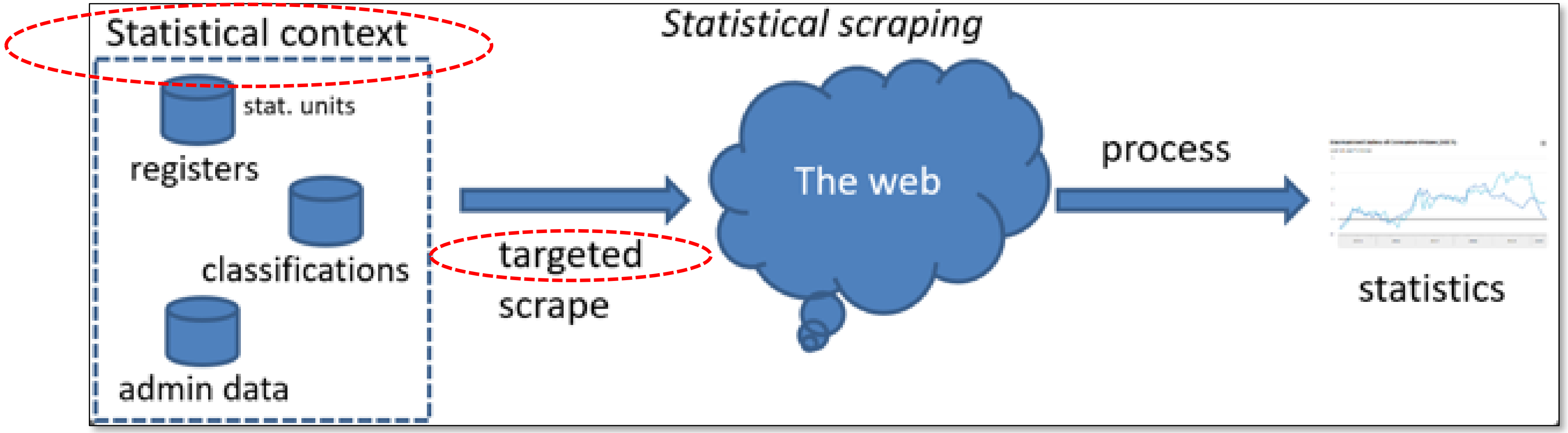
Pseudo-design
based estimate

This is a convenience
sample, use it as a fast beta
indicator if strong
correlation is proven.

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

Definition

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.*



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
- Search (engine) ***leakage***: needs attention bot ***manageable***

Wrap-up

- ***Online data*** have shown to be valuable for official statistics, either as primary or auxiliary data source
- In addition to bulk scraping a ***statistical scraping*** approach has been presented where a-priori information is used for a ***targeted scrape*** resulting in selected samples with high statistical value
- Examples can be found in in ***business register enhancement*** and ***price***, ***tourism*** and ***education*** statistics. Applicability in other domains is to be explored.
- Statistical scraping has methodological advantages, to cope with representation errors and if applied on unit level proven ***quality indicators*** can be calculated
- Statistical scraping leads to ***smaller*** and more manageable data streams
- Attention should be paid to treatment of possible ***sensitive*** scraping input data
- The concept may ***complement*** or in some cases ***replace*** bulk scraping methods

Thanks

Questions, comments, ideas:

Olav ten Bosch
o.tenbosch@cbs.nl

And please like the awesome list of official statistics software:

awesomeofficialstatistics.org

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union