Business register improvements: a balance between search, scrape and 3rd party web data

Statistics Netherlands

Olav ten Bosch, Arnout van Delden, Nick de Wolf

NTTS 2023, Brussels, 6-9 March 2023

Contents

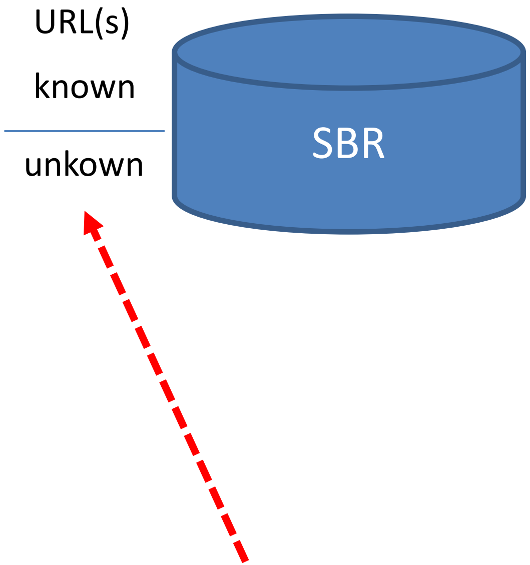
- The main concept
- Search, scrape and linking 3rd party web data
- Other data sources
- Putting it all together
- Zooming out: web data and survey design
- Wrap up

The main concept

SBR

The web

The main concept



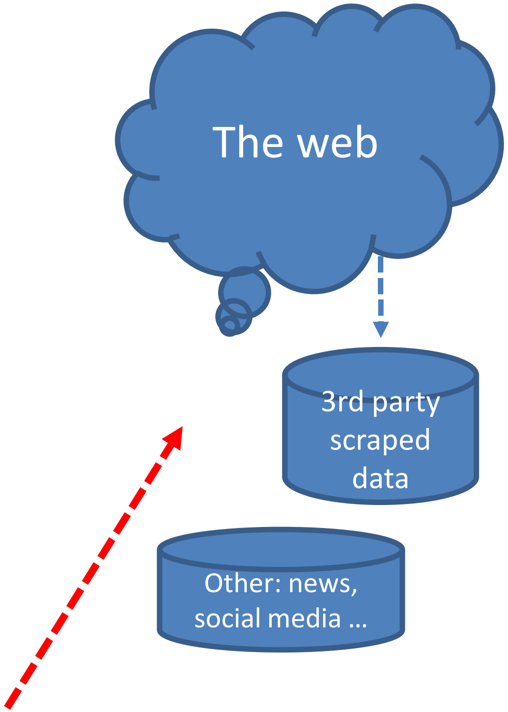
URL(s)

known

unkown

SBR

The ratio known/unknown is country- specific



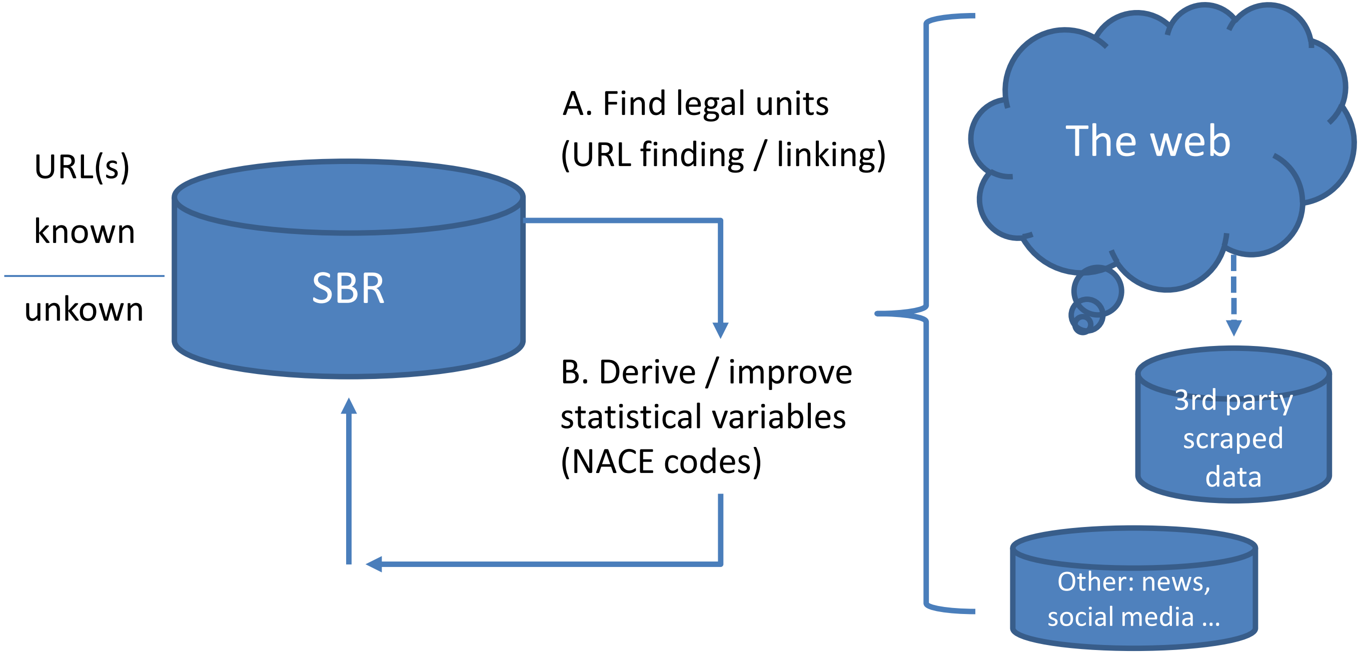
The web

3rd party scraped data

Other: news, social media …

Web data is more than websites only

The main concept



URL(s)

known

unkown

SBR

A. Find legal units
(URL finding / linking)

B. Derive / improve
statistical variables
(NACE codes)

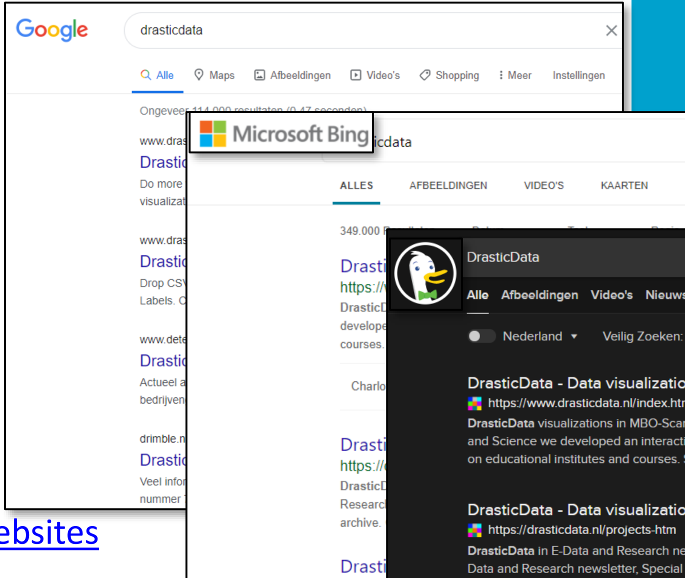
The web

3rd party
scraped
data

Other: news,
social media …

Search (1)

- To ***find*** or ***verify*** URL’s for legal units (LUs)
- ***Automatically*** search on what we know from SBR
  - e.g. name, address, municipality, id, and/or contact info
- Using free or paid ***API***
- Search engine ***leakage manageable:***
  - Use paid/trusted search engines
  - Use search phrase wisely
  - Spread across search engines and in time



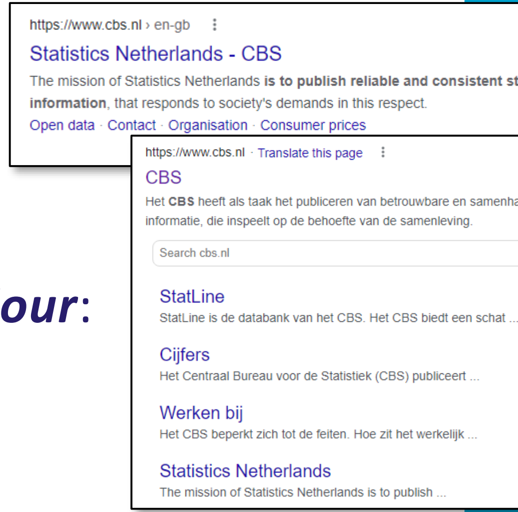
DrasticData - Data visualizatio
https://www.drasticdata.nl/index.htm
DrasticData visualizations in MBO-Scan
and Science we developed an interacti
on educational institutes and courses.

DrasticData - Data visualizatio
https://drasticdata.nl/projects-htm
DrasticData in E-Data and Research ne
Data and Research newsletter, Special

https://ec.europa.eu/eurostat/cros/content/url-finding-methodology_en
https://www.cbs.nl/en-gb/background/2020/01/searching-for-business-websites

Search (2)

Snippet examples



https://www.cbs.nl › en-gb

Statistics Netherlands - CBS

The mission of Statistics Netherlands **is to publish reliable and consistent st** **information**, that responds to society's demands in this respect.

Open data · Contact · Organisation · Consumer prices

https://www.cbs.nl · Translate this page

CBS

Het **CBS** heeft als taak het publiceren van betrouwbare en samenha informatie, die inspeelt op de behoefte van de samenleving.

Search cbs.nl

StatLine

StatLine is de databank van het CBS. Het CBS biedt een schat ...

Cijfers

Het Centraal Bureau voor de Statistiek (CBS) publiceert ...

Werken bij

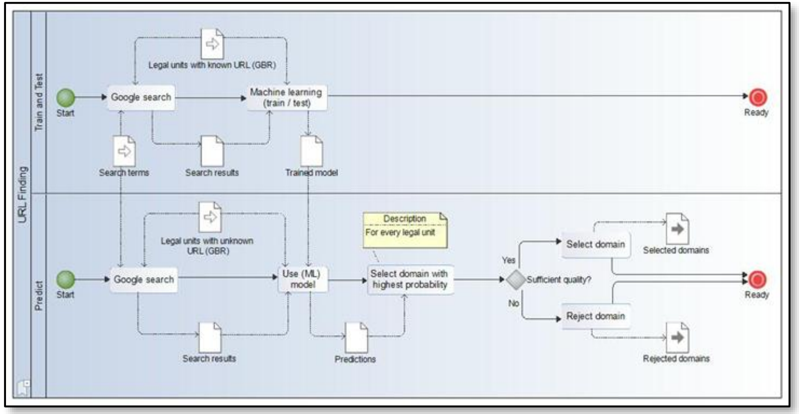
Het CBS beperkt zich tot de feiten. Hoe zit het werkelijk ...

Statistics Netherlands

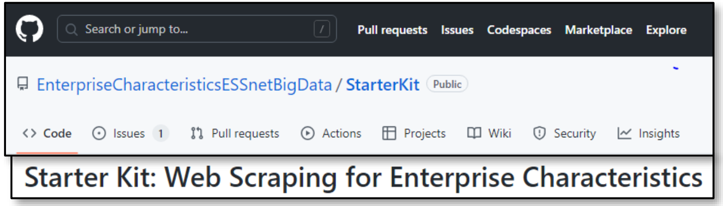
The mission of Statistics Netherlands is to publish ...

How to select the right URL from list of search results?

- Using ***snippet*** and/or ***extra scraping*** step
- Use an ***ML*** model capturing the ***search engine behaviour***:
  - Train and test on set of LUs with known URL
  - Predict URL from list of search results



https://github.com/SNStatComp/urlfinding



EnterpriseCharacteristicsESSnetBigData / StarterKit

Starter Kit: Web Scraping for Enterprise Characteristics

https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit

Scrape (1)

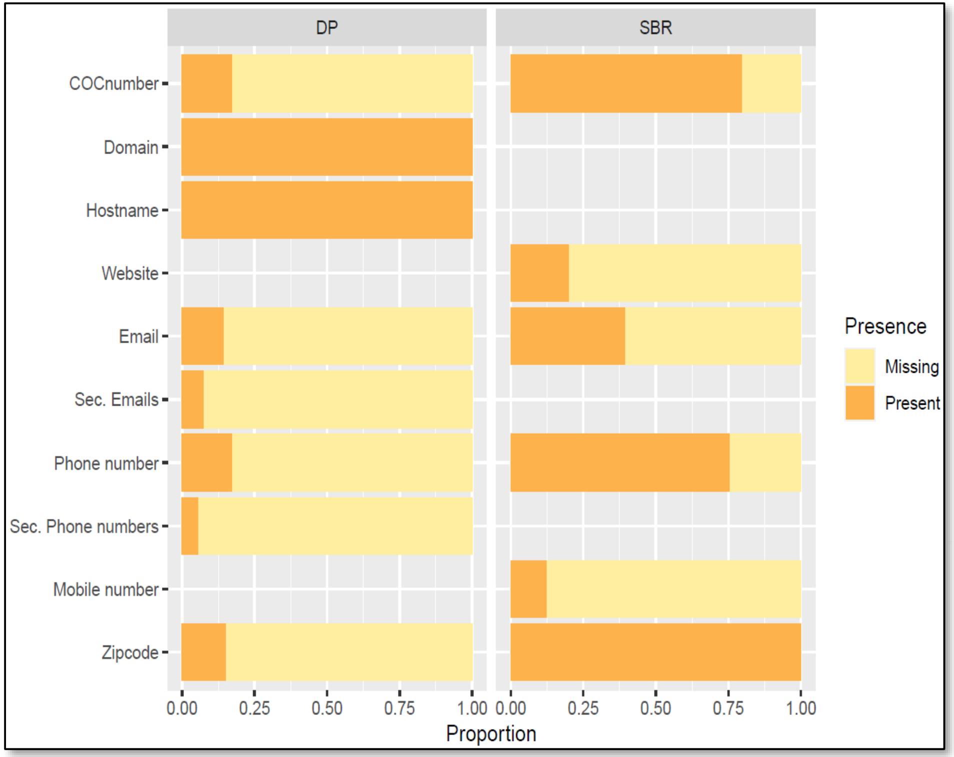
- Two types of scraping:
  - ***Generic***: no prior knowledge of site structure
  - ***Specific***: scraper is designed for specifics of website
- For SBR enhancement:
  - ***generic*** scraping, usually scraping website up to a certain ***depth***
- What to ***store***:
  - Complete website, only texts or variables derived?
- ***Focused*** scraper:
  - gives priority to those parts of websites that are expected to contain valuable info, for example “about us” or “vacancies”

Scrape (2)

- National legislation might enforce ***identifying information*** on enterprise website
  - tax-id or COC-id
  - this is profitable in scraping for official statistics!
- Be aware of ***n-to-m relationships*** LU <-> website
  - LU might have multiple websites
  - Register the main website (if identified) or all?
  - Small business might be present only on business services portal listing many different small companies

Linking 3rd party web data (1)

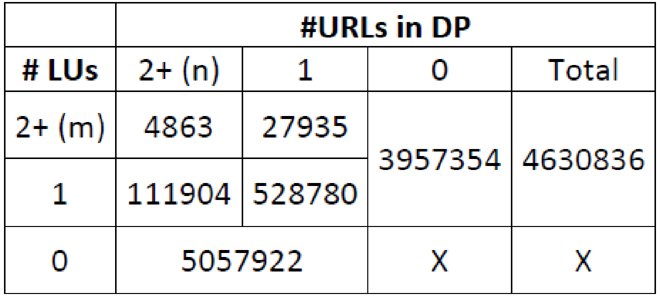
- Use web data collected by 3rd parties if added value is considerable
- Statistics Netherlands:
  > 2 yrs experience
  DataProvider (DP) data
- Monthly datasets
- Gaps complicate linking:



Linking 3rd party web data (2)

- n-to-m relationships DP<->SBR
- In our case: 11% 1-to-1; 2.5% n-to-1

Table 1. Nr. of legal units (LUs) by linkage cardinality at 75% linkage probability



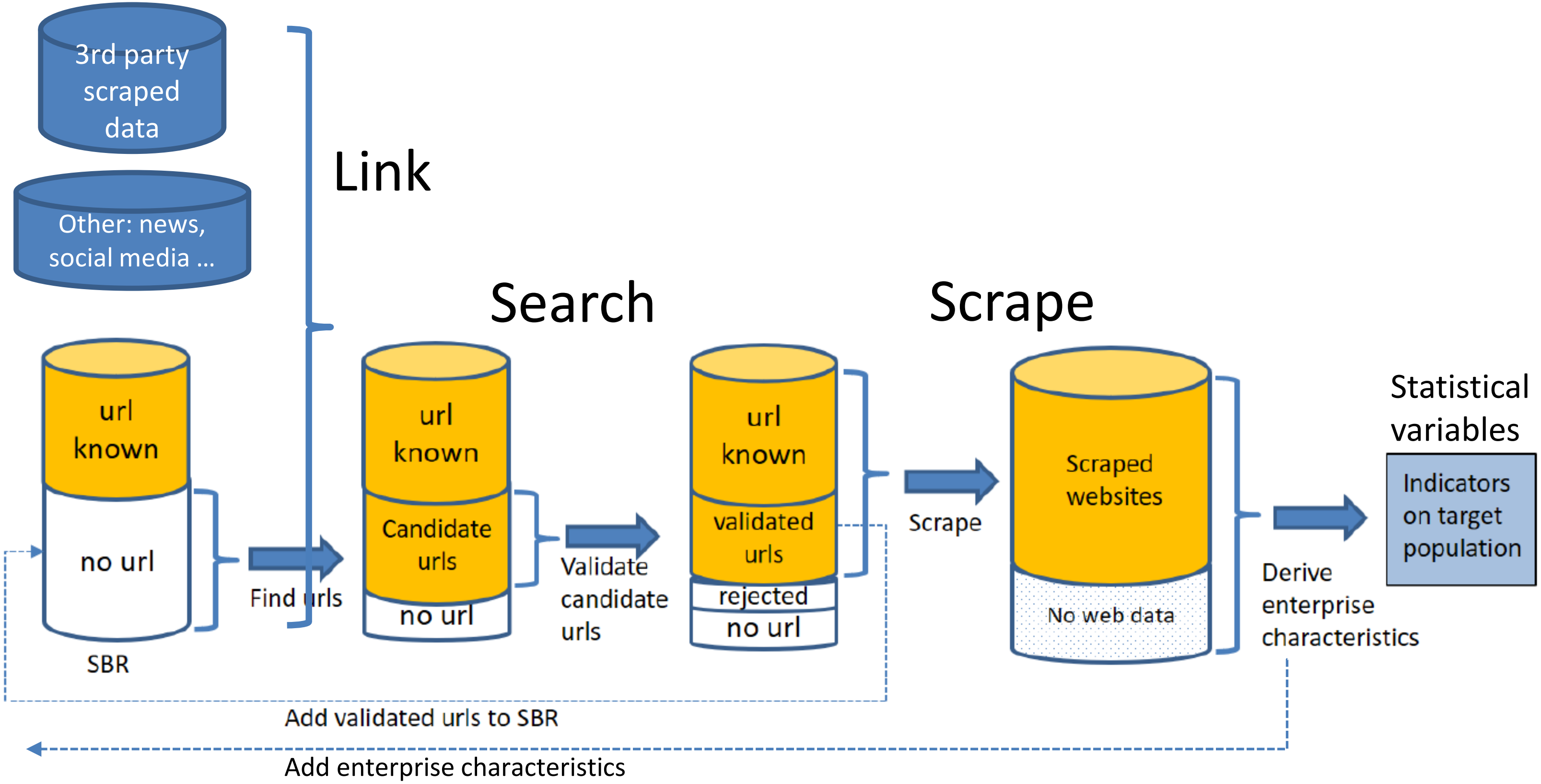
	#URLs in DP			
# LUs	2+ (n)	1	0	Total
2+ (m)	4863	27935	3957354	4630836
1	111904	528780		
0	5057922		X	X

- For 14.5% LUs a URL could be deduced from 3rd party web data => using 3rd party web data makes sense!
- Linking strategy still being refined

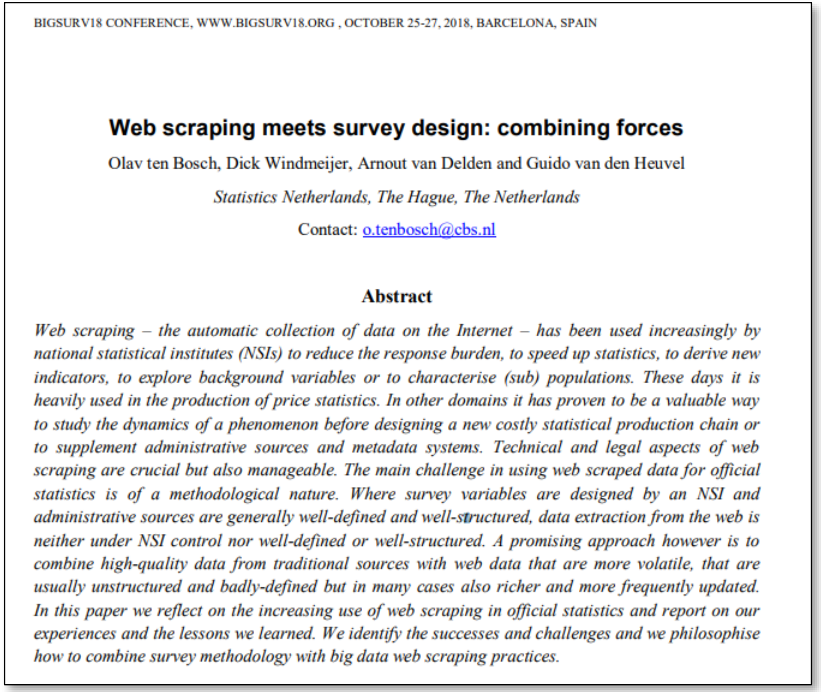
Other web data sources

- Use ***domain registry*** to deduce URLs:
  - degree of openness varies per country and domain
  - .nl domain is not open by default
- ***Press releases***, ***social media***
- ***Online Job ads*** (OJAs)
  - Can we use OJAs to improve our knowledge about economic activity of a LU?
  - Linking challenge: OJA <-> LU

Putting it all together



Zooming out: web data and survey design (1)



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

Web scraping meets survey design: combining forces

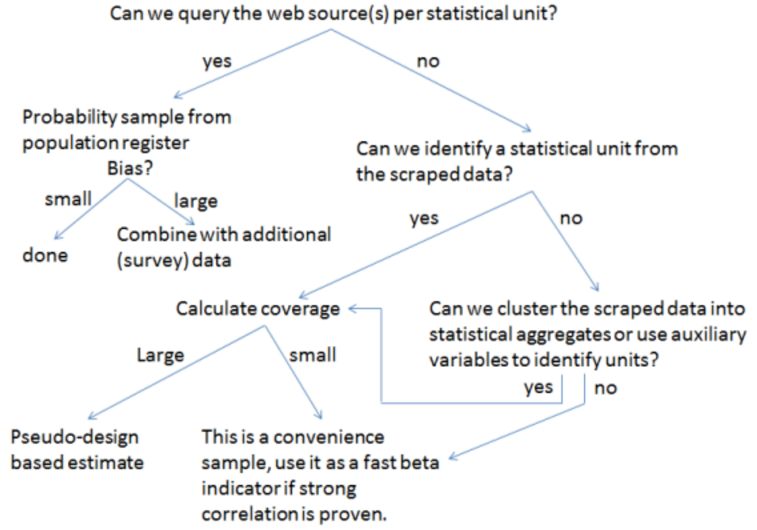
Olav ten Bosch, Dick Windmeijer, Arnout van Delden and Guido van den Heuvel

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.



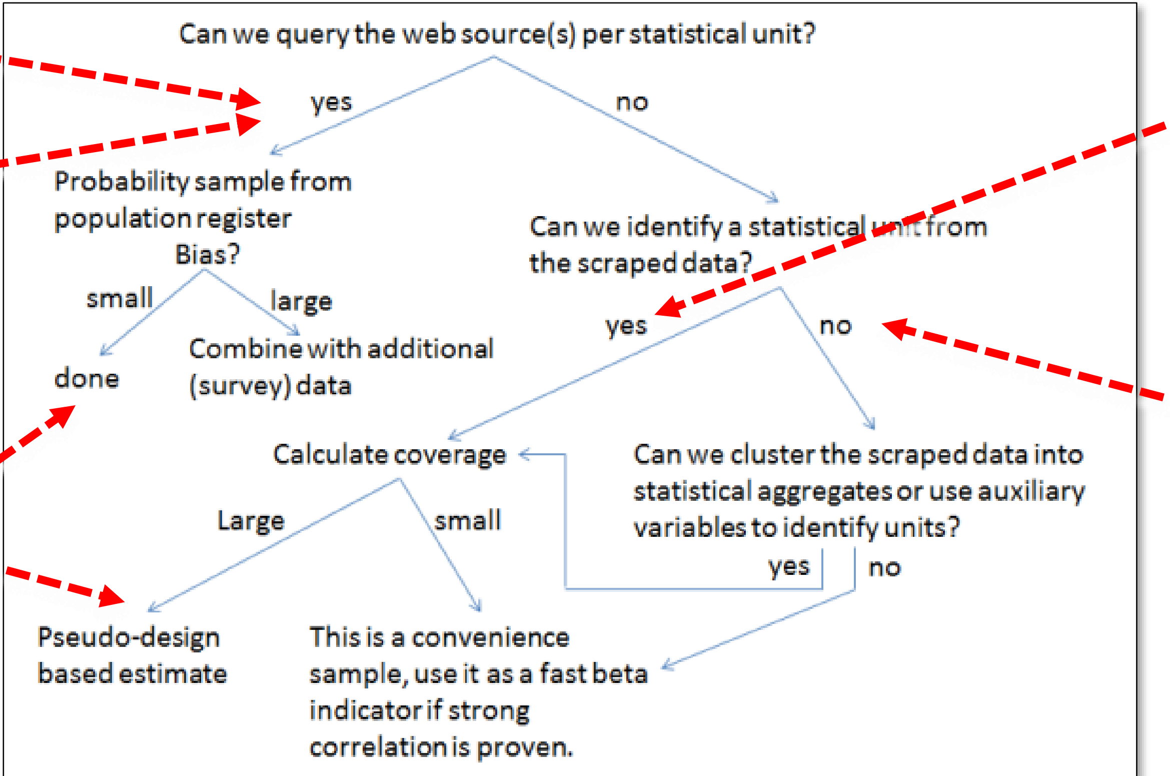
Can we query the web source(s) per statistical unit?
yes
no
Probability sample from population register
Bias?
small
large
done
Combine with additional (survey) data
Can we identify a statistical unit from the scraped data?
yes
no
Calculate coverage
Can we cluster the scraped data into statistical aggregates or use auxiliary variables to identify units?
Large
small
yes
no
Pseudo-design based estimate
This is a convenience sample, use it as a fast beta indicator if strong correlation is proven.

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

General workflow for any web source

BigSurv2018
https://www.researchgate.net/publication/327385487_Web_scraping_meets_survey_design_combining_forces

Zooming out: web data and survey design (2)



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

Large

small

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

yes

no

Pseudo-design based estimate

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

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

URL finding

Scraping a LU with known URL

Deriving statistical variables

Linked web data

Unlinked web data

Wrap up

- SBR enhancements from web data: mix of ***search***, ***scraping*** and linking ***3rd party web data***
- Search ***methodology*** is ready; ***leakage*** is manageable
- Scraping: ***generic***, ***focussed***, use ***identifying*** information
- Linking 3rd party data: proven to be ***valuable***
- In all cases: ***n-to-m relationships*** LU <-> web data
- Optional: domain registry, news, social media, OJAs
- More general view: ***web data*** and ***survey design***

Questions, ideas, suggestions

?

Olav ten Bosch
o.tenbosch@cbs.nl

and please keep an eye on the awesome list of open source software:
awesomeofficialstatistics.org