Business register quality enhancement

WP3-UC5 members:
CBS, HSL, SCB, SF, STATA

WIH-CON, June 12 2023, Brussels

Trusted Smart Statistics – Web Intelligence Network
Grant Agreement: 101035829

Web Intelligence
Network

Funded by
the European Union

Contents

- The main concept
- Search, scrape, link, train, predict, derive
- Other data sources
- Putting it all together
- Zooming out: web data and survey design
- Future work
- Wrap up

Funded by the European Union

The main concept (1)

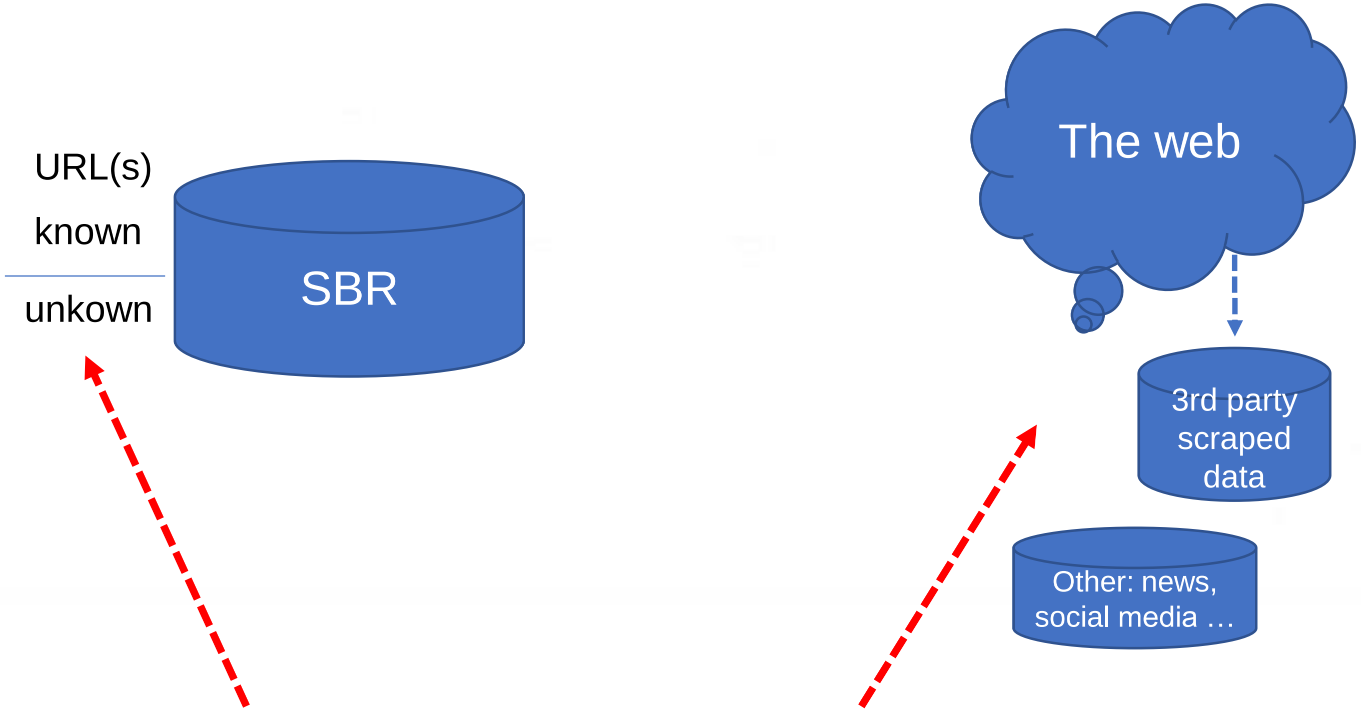
SBR

Statistical Business Register

The web

Funded by the European Union

The main concept (2)



URL(s)

known

unkown

SBR

The web

3rd party scraped data

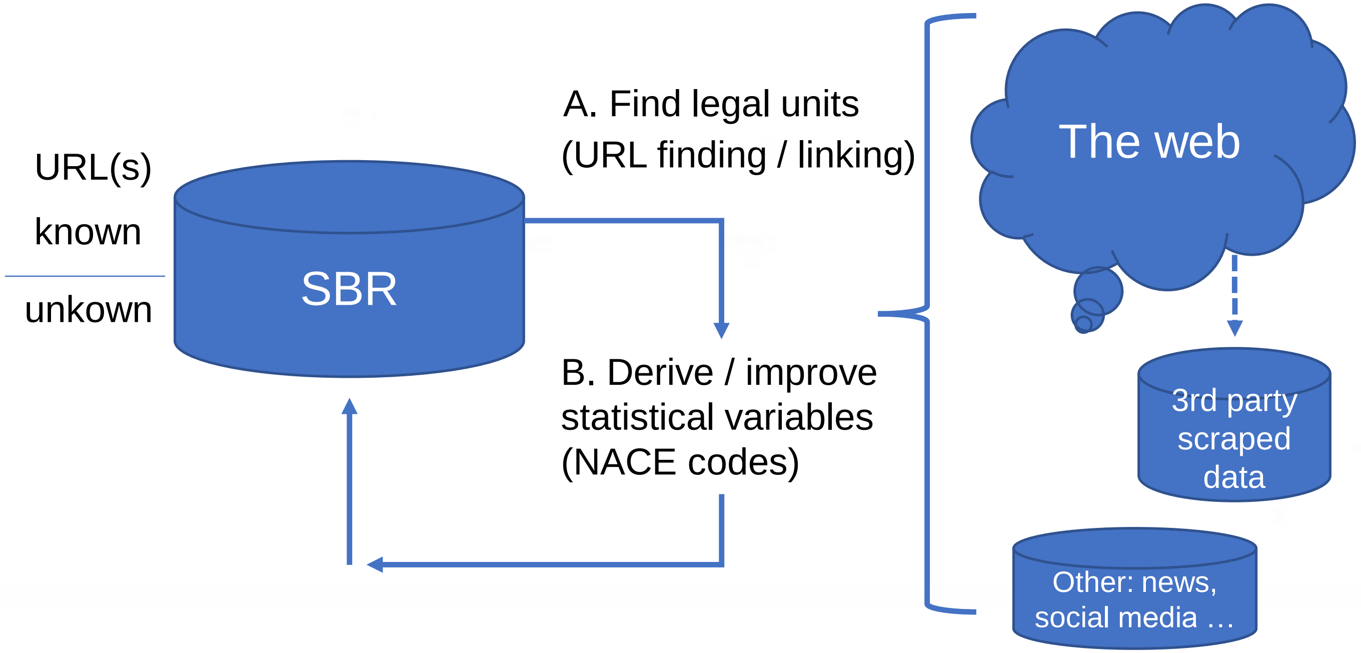
Other: news, social media …

The ratio known/unknown is country-specific

Web data is more than websites only

Funded by the European Union

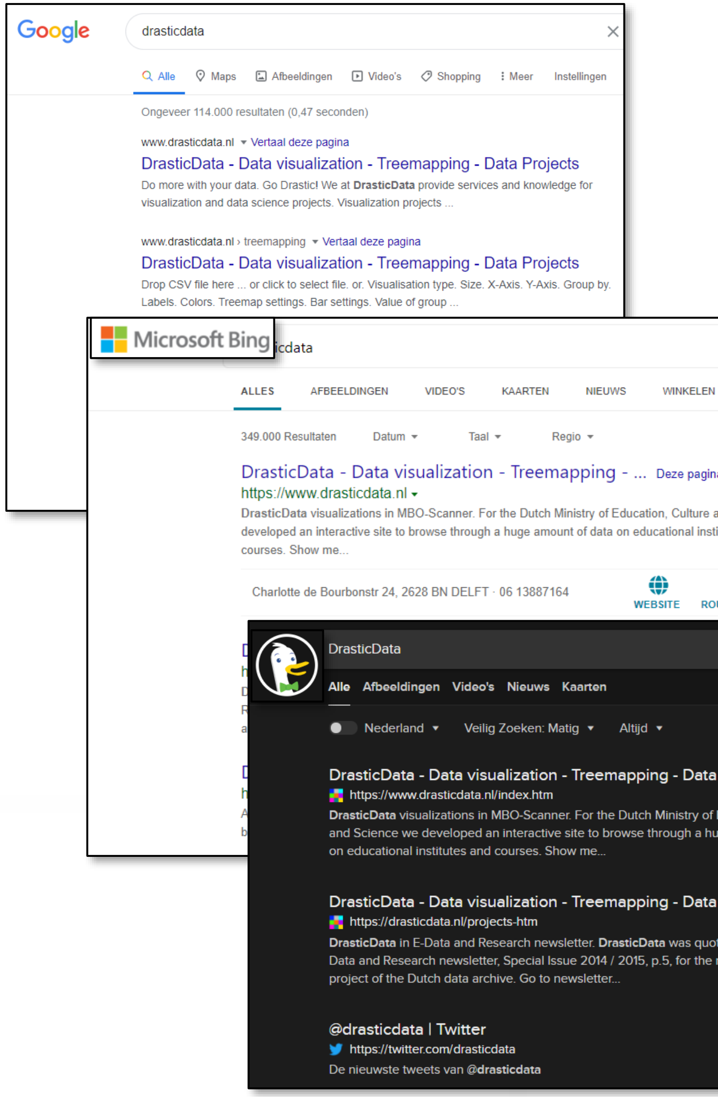
The main concept (3)



Search (1)

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

https://ec.europa.eu/eurostat/cros/content/url-finding-methodology_en



Ongeveer 114.000 resultaten (0,47 seconden)

www.drasticdata.nl ▾ Vertaal deze pagina
DrasticData - Data visualization - Treemapping - Data Projects
Do more with your data. Go Drastic! We at **DrasticData** provide services and knowledge for visualization and data science projects. Visualization projects ...

www.drasticdata.nl › treemapping ▾ Vertaal deze pagina
DrasticData - Data visualization - Treemapping - Data Projects
Drop CSV file here ... or click to select file. or. Visualisation type. Size. X-Axis. Y-Axis. Group by. Labels. Colors. Treemap settings. Bar settings. Value of group ...

349.000 Resultaten Datum ▾ Taal ▾ Regio ▾

DrasticData - Data visualization - Treemapping - ... Deze pagina
https://www.drasticdata.nl ▾
DrasticData visualizations in MBO-Scanner. For the Dutch Ministry of Education, Culture a
developed an interactive site to browse through a huge amount of data on educational insti
courses. Show me...

Charlotte de Bourbonstr 24, 2628 BN DELFT · 06 13887164

Nederland ▾ Veilig Zoeken: Matig ▾ Altijd ▾

DrasticData - Data visualization - Treemapping - Data
https://www.drasticdata.nl/index.htm
DrasticData visualizations in MBO-Scanner. For the Dutch Ministry of I
and Science we developed an interactive site to browse through a hu
on educational institutes and courses. Show me...

DrasticData - Data visualization - Treemapping - Data
https://drasticdata.nl/projects-htm
DrasticData in E-Data and Research newsletter. **DrasticData** was quot
Data and Research newsletter, Special Issue 2014 / 2015, p.5, for the
project of the Dutch data archive. Go to newsletter...

@drasticdata | Twitter
https://twitter.com/drasticdata
De nieuwste tweets van **@drasticdata**

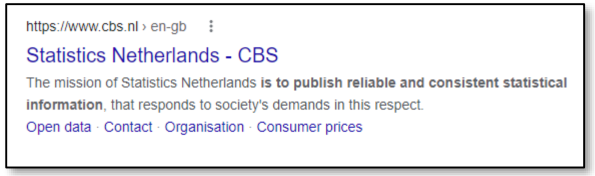
Funded by the European Union

Search (2)

How to select the right URL from a 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

Snippet examples

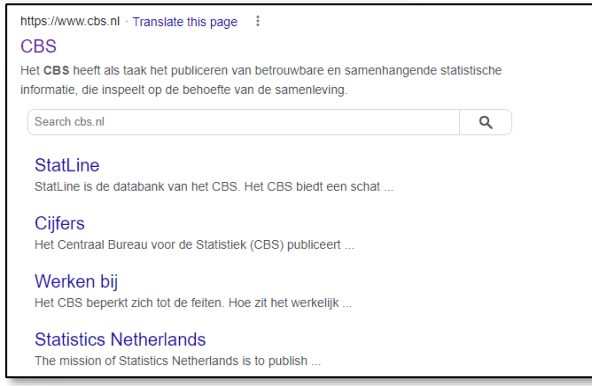


https://www.cbs.nl › en-gb

Statistics Netherlands - CBS

The mission of Statistics Netherlands **is to publish reliable and consistent statistical 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 samenhangende statistische informatie, die inspeelt op de behoefte van de samenleving.

StatLine

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

Cijfers

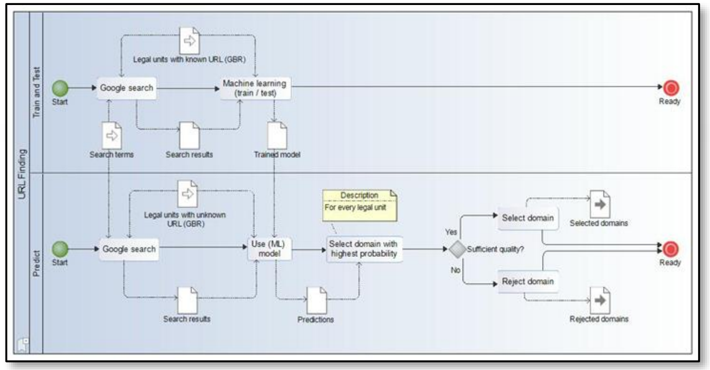
Het Centraal Bureau voor de Statistiek (CBS) publiceert ...

Werken bij

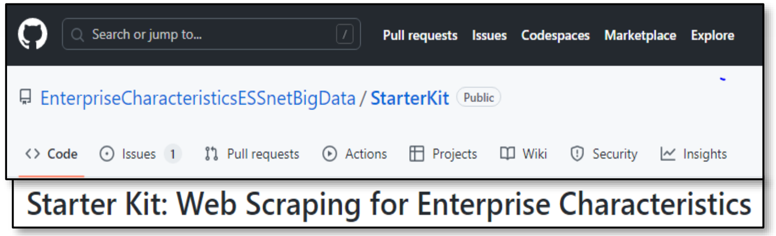
Het CBS beperkt zich tot de feiten. Hoe zit het werkelijk ...

Statistics Netherlands

The mission of Statistics Netherlands is to publish ...



https://github.com/SNStatComp/urlfinding



Starter Kit: Web Scraping for Enterprise Characteristics

https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit

Web Intelligence Network

Funded by the European Union

Scrape

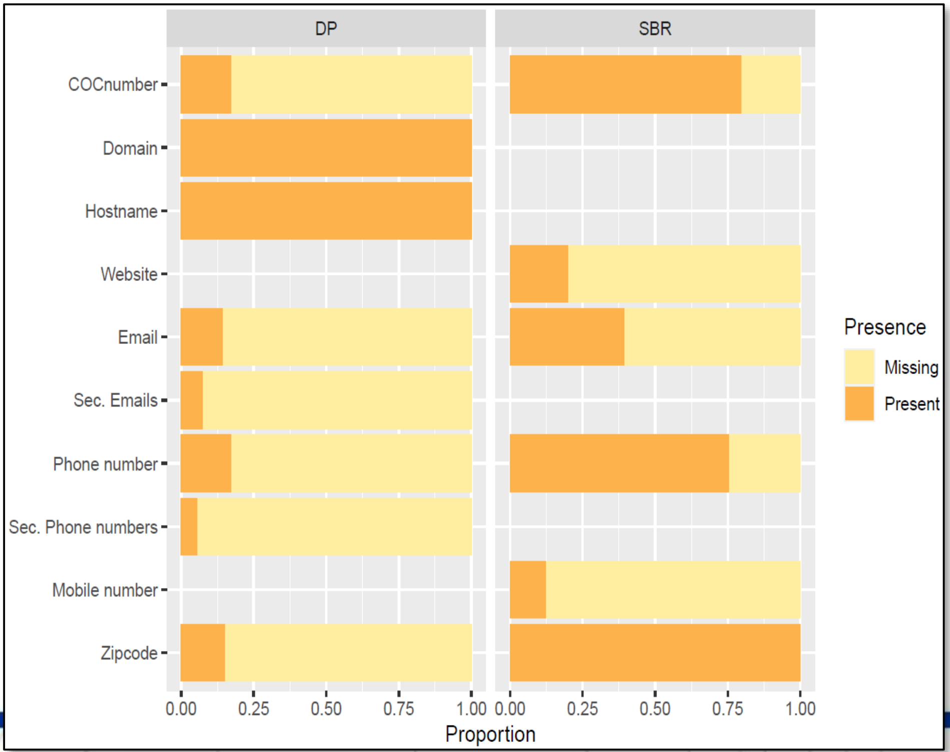
- Two types:
  - ***Generic***: no prior knowledge of site structure
  - ***Specific***: scraper is designed for specifics of website
- ***Focused*** scraper:
  - Gives priority to those parts of websites that are expected to contain valuable info, for example “about us” or “vacancies”
- National legislation might enforce ***identifying information*** on websites:
  - tax-id or COC-id
- 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

Web Intelligence Network

Funded by the European Union

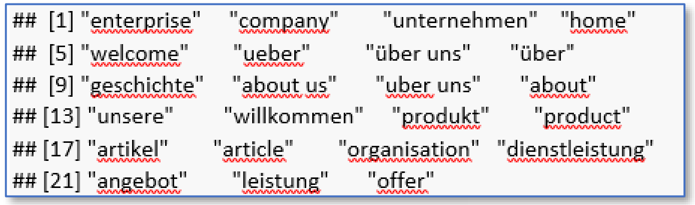
Link 3rd party web data

- Use web data collected by 3rd parties if added value is considerable
- NL > 2 yrs experience with DataProvider (DP) data
- Monthly datasets, linking to SBR
- Not always easy, gaps in data
- For ~15% of LUs a URL could be deduced



Web Intelligence Network
Funded by the European Union

Train, predict, derive



```
##  [1] "enterprise"    "company"      "unternehmen"   "home"
##  [5] "welcome"      "ueber"        "über uns"      "über"
##  [9] "geschichte"   "about us"     "uber uns"      "about"
## [13] "unsere"       "willkommen"   "produkt"       "product"
## [17] "artikel"      "article"      "organisation"  "dienstleistung"
## [21] "angebot"      "leistung"     "offer"
```

NACE detection

- AT: Word-driven NACE-1 prediction (XGBoost)
- NL: Predict whether a registered NACE is incorrect
- SE: NACE detection experiments with KB-BERT method adapted and extended for Swedish language

Correcting or complement administrative information:

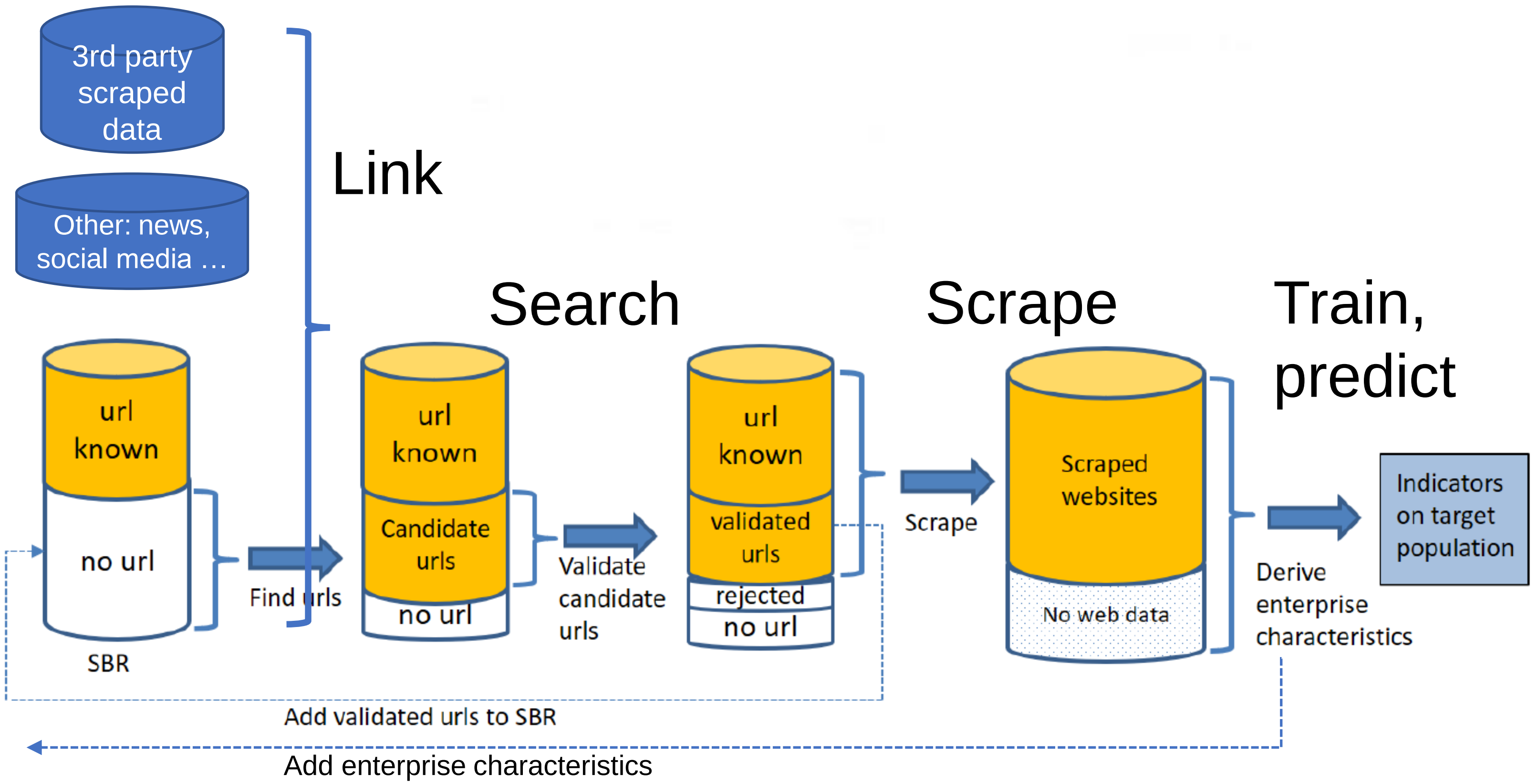
- HE: contact information discovery from websites
  - emails, classified into functional/high/medium/low

Other data sources

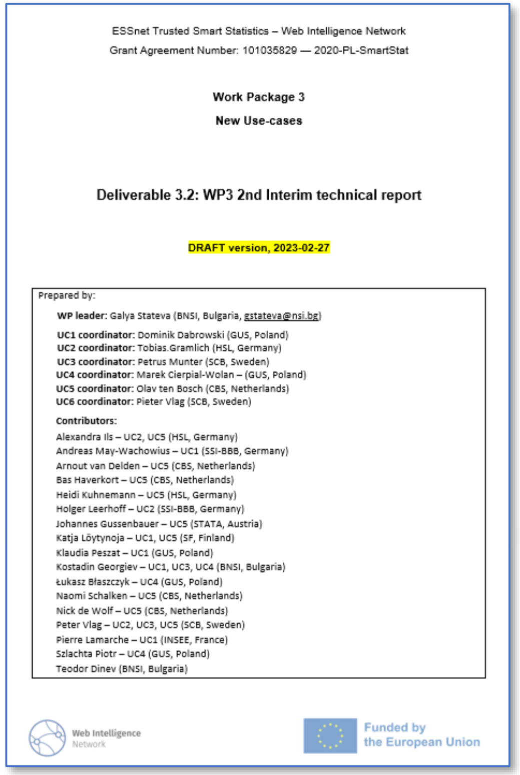
- Use domain registry for (additional) URL finding:
  - Degree of openness varies per country and domain
  - .nl domain is not open; .fi domain has public API
- Wikipedia / dbpedia / business collection portals
- News releases, social media
- Financial / annual reports
- OJAs

Funded by the European Union

All together



All together: work in year 2 of project:



ESSnet Trusted Smart Statistics – Web Intelligence Network
Grant Agreement Number: 101035829 — 2020-PL-SmartStat

Work Package 3
New Use-cases

Deliverable 3.2: WP3 2nd Interim technical report

DRAFT version, 2023-02-27

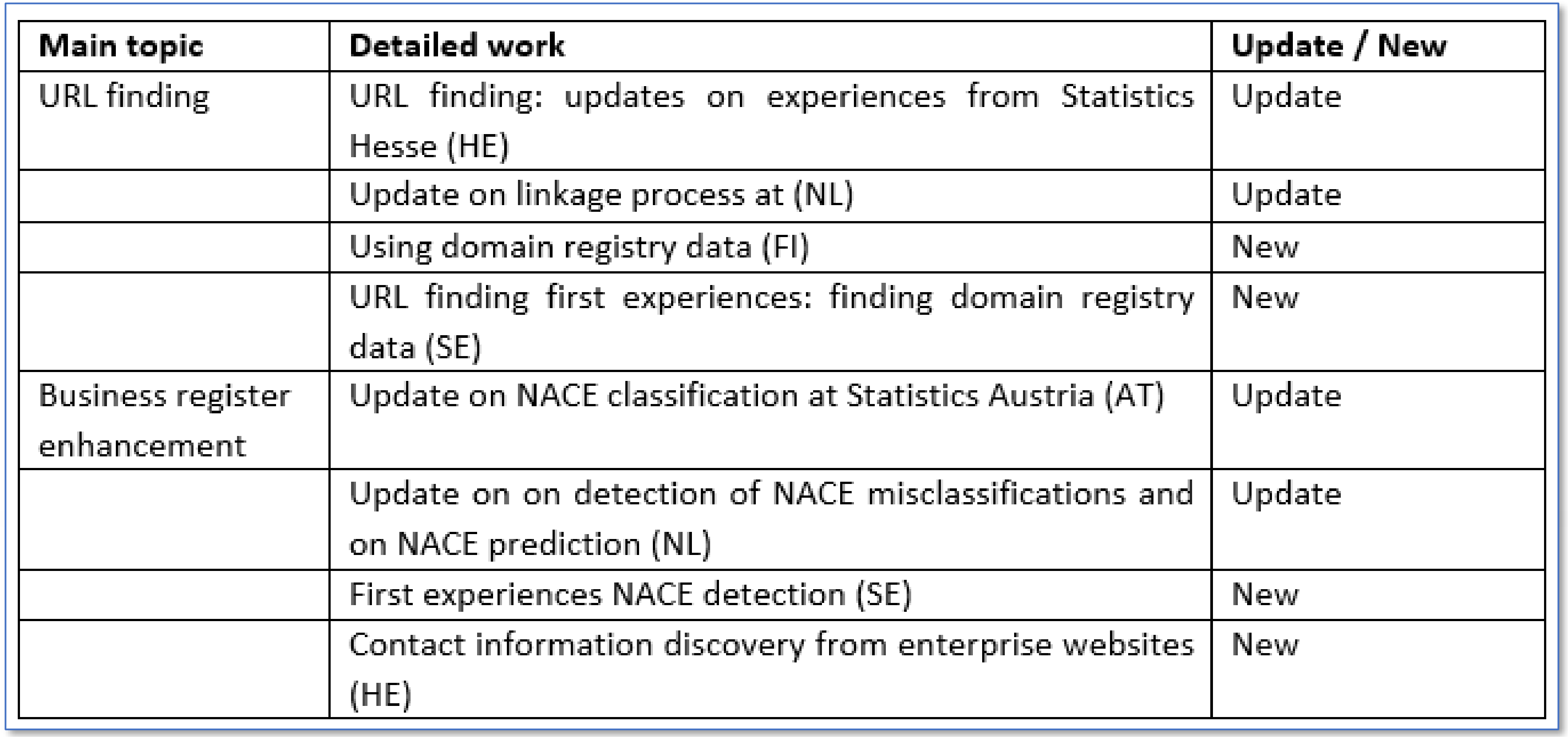
Prepared by:

WP leader: Galya Stateva (BNSI, Bulgaria, gstateva@nsi.bg)

UC1 coordinator: Dominik Dabrowski (GUS, Poland)
UC2 coordinator: Tobias.Gramlich (HSL, Germany)
UC3 coordinator: Petrus Munter (SCB, Sweden)
UC4 coordinator: Marek Cierpial-Wolan – (GUS, Poland)
UC5 coordinator: Olav ten Bosch (CBS, Netherlands)
UC6 coordinator: Pieter Vlag (SCB, Sweden)

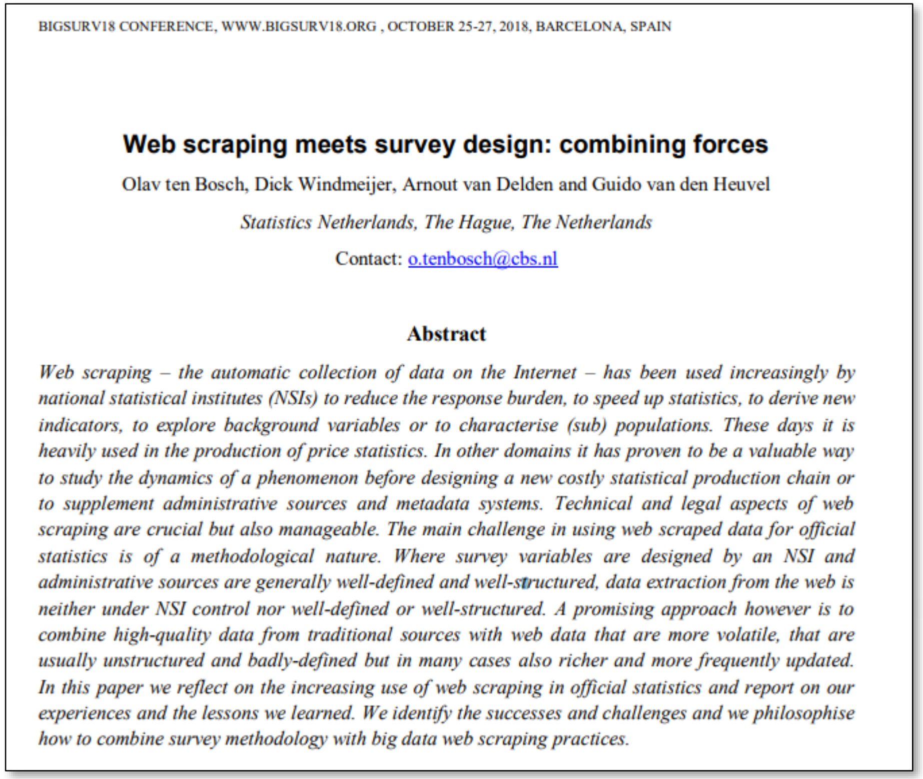
Contributors:
Alexandra Ils – UC2, UC5 (HSL, Germany)
Andreas May-Wachowius – UC1 (SSI-BBB, Germany)
Arnout van Delden – UC5 (CBS, Netherlands)
Bas Haverkort – UC5 (CBS, Netherlands)
Heidi Kuhnemann – UC5 (HSL, Germany)
Holger Leerhoff – UC2 (SSI-BBB, Germany)
Johannes Gussenbauer – UC5 (STATA, Austria)
Katja Löytynoja – UC1, UC5 (SF, Finland)
Klaudia Peszat – UC1 (GUS, Poland)
Kostadin Georgiev – UC1, UC3, UC4 (BNSI, Bulgaria)
Łukasz Błaszczyk – UC4 (GUS, Poland)
Naomi Schalken – UC5 (CBS, Netherlands)
Nick de Wolf – UC5 (CBS, Netherlands)
Peter Vlag – UC2, UC3, UC5 (SCB, Sweden)
Pierre Lamarche – UC1 (INSEE, France)
Szlachta Piotr – UC4 (GUS, Poland)
Teodor Dinev (BNSI, Bulgaria)

Deliverable 3.2: WP3 2nd Interim technical report



Main topic	**Detailed work**	**Update / New**
URL finding	URL finding: updates on experiences from Statistics Hesse (HE)	Update
	Update on linkage process at (NL)	Update
	Using domain registry data (FI)	New
	URL finding first experiences: finding domain registry data (SE)	New
Business register enhancement	Update on NACE classification at Statistics Austria (AT)	Update
	Update on on detection of NACE misclassifications and on NACE prediction (NL)	Update
	First experiences NACE detection (SE)	New
	Contact information discovery from enterprise websites (HE)	New

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.

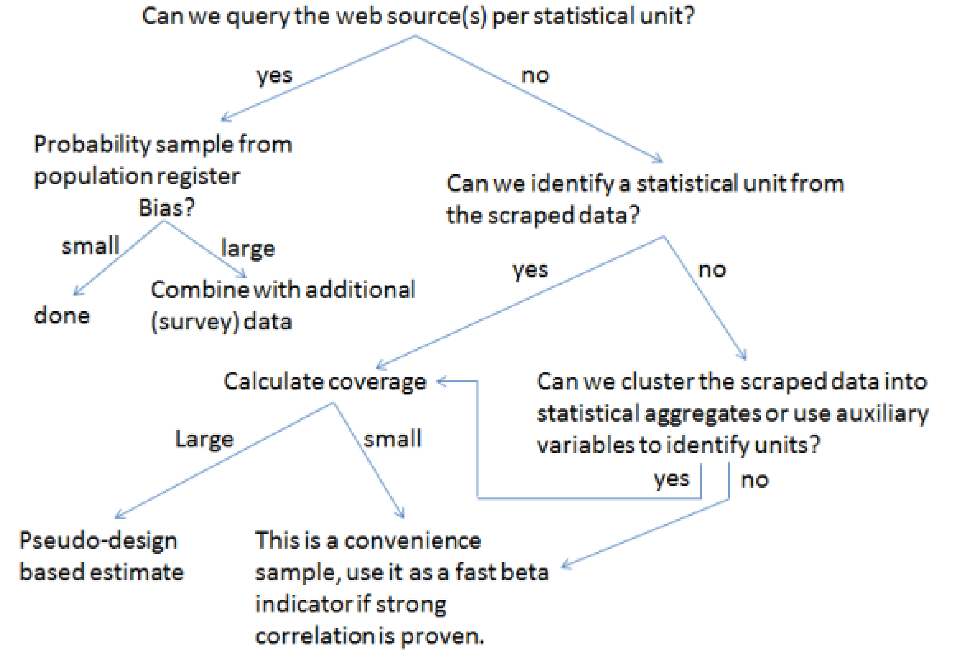

Can we query the web source(s) per statistical unit?
yes
no
Probability sample from population register
Bias?
small
large
done
Combine with additional (survey) data
Can we identify a statistical unit from the scraped data?
yes
no
Calculate coverage
Large
small
Pseudo-design based estimate
This is a convenience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

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

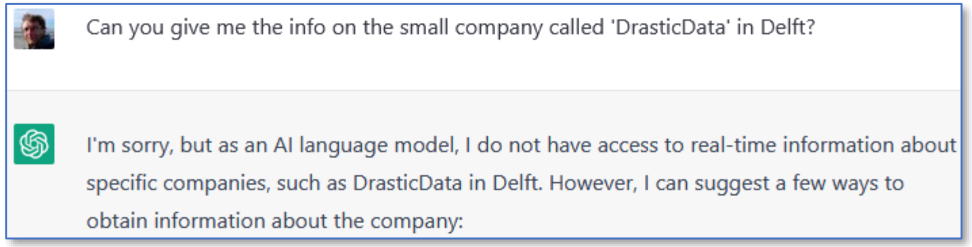
General workflow for any web source

BigSurv2018

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

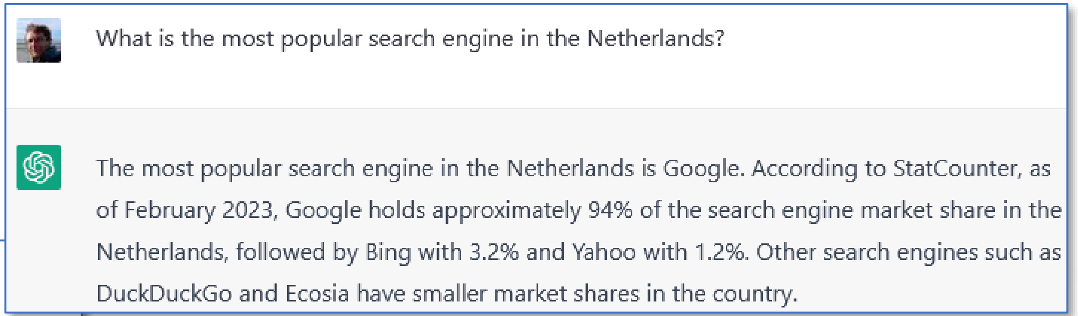
Future

- What can web data tell us about ***enterprise networks***?
- Can ***AI***, such as Chat-GPT help us finding websites or deriving variables?



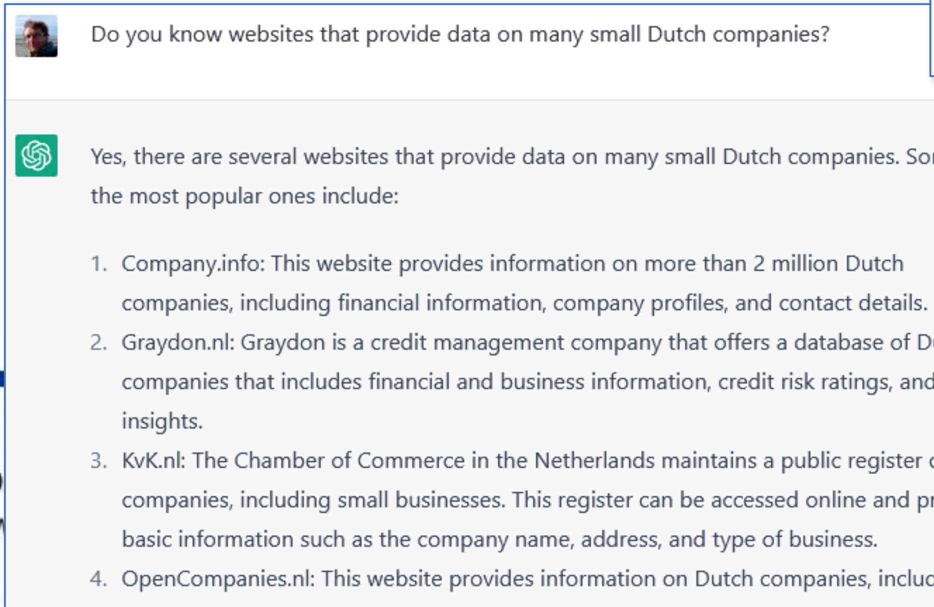
Can you give me the info on the small company called 'DrasticData' in Delft?

I'm sorry, but as an AI language model, I do not have access to real-time information about specific companies, such as DrasticData in Delft. However, I can suggest a few ways to obtain information about the company:



What is the most popular search engine in the Netherlands?

The most popular search engine in the Netherlands is Google. According to StatCounter, as of February 2023, Google holds approximately 94% of the search engine market share in the Netherlands, followed by Bing with 3.2% and Yahoo with 1.2%. Other search engines such as DuckDuckGo and Ecosia have smaller market shares in the country.



Do you know websites that provide data on many small Dutch companies?

Yes, there are several websites that provide data on many small Dutch companies. Some of the most popular ones include:

1. Company.info: This website provides information on more than 2 million Dutch companies, including financial information, company profiles, and contact details.
2. Graydon.nl: Graydon is a credit management company that offers a database of Dutch companies that includes financial and business information, credit risk ratings, and other insights.
3. KvK.nl: The Chamber of Commerce in the Netherlands maintains a public register of companies, including small businesses. This register can be accessed online and provides basic information such as the company name, address, and type of business.
4. OpenCompanies.nl: This website provides information on Dutch companies, including

Funded by
the European Union

Wrap up

- The web is a rich source on enterprise information for official statistics
- Starting from a Statistical Business Register we use a mix of searching, scraping, linking 3rd party web data and machine learning to enhance the business register
- Be aware of n-to-m relationships LU <-> web data
- Other sources can be domain registry, news, social media, OJAs
- This fits into a more general view: web data and survey design
- Future topics: enterprise networks, AI

Funded by the European Union

Questions / ideas welcome

- Olav ten Bosch (CBS, Netherlands), coordinator, o.tenbosch@cbs.nl
- Alexandra Ils (HSL, Germany)
- Arnout van Delden (CBS, Netherlands)
- Heidi Kuhnemann (HSL, Germany)
- Johannes Gussenbauer (STATA, Austria)
- Katja Löytynoja (SF, Finland)
- Nick de Wolf (CBS, Netherlands)
- Pieter Vlag (SCB, Sweden)
- and others

Funded by the European Union