Statistics Poland

Current Challenges and Possible Solutions for the Use of Web Data as a Source for Official Statistics

Jacek Maślankowski, University of Gdańsk, Statistics Poland

Piet Daas, Technical University of Eindhoven, Statistics Netherlands

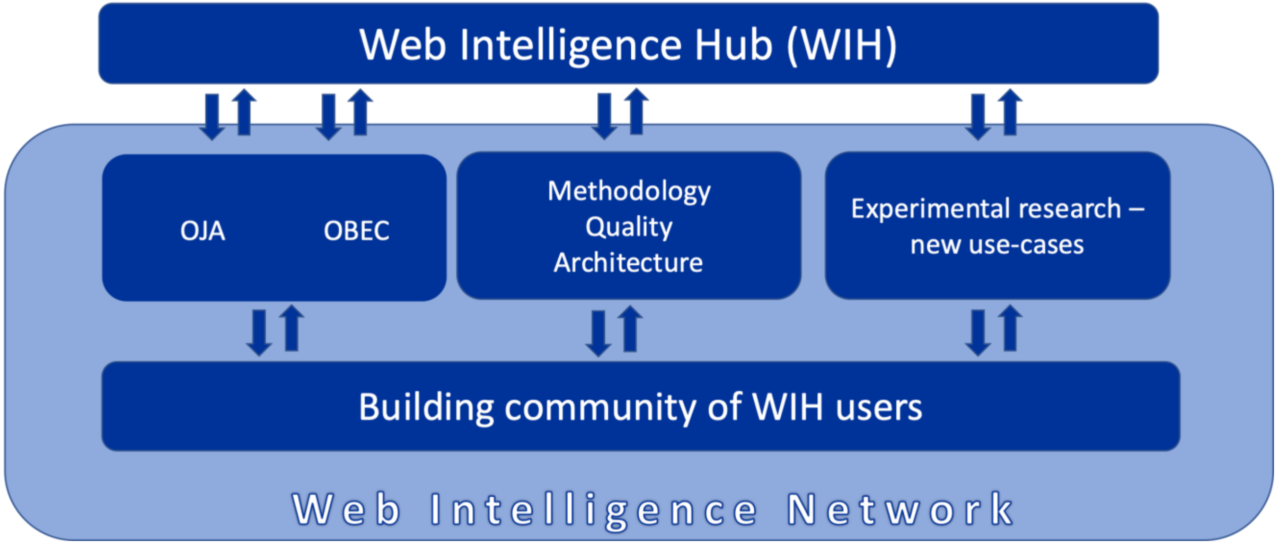
MET2023

Agenda

- Prerequisites
  - Gap identified
  - Web Intelligence
  - Population perspective
  - URLs perspective
- Research method and data collection
- Case study
- Results
- Conclusions and final remarks

Prerequisites – gap identified

- Web Intelligence Network (NSIs)
- Web Intelligence Hub (Eurostat)
  - Web Intelligence Platform (Eurostat)
  - Web Inteligence Datalab (Eurostat)



Web Intelligence Hub (WIH)
OJA
OBEC
Methodology
Quality
Architecture
Experimental research – new use-cases
Building community of WIH users
Web Intelligence Network

Source: Peszat K., Maślankowski J., How WIN supports the WIH, DIME 2023

Prerequisites – Web Intelligence

- Web Intelligence results in number of methodological challenges
- The majority of them are generic for all scraping processes and need to be dealt with as good as possible („learning by doing”)
- Web scraping is a method to get the source of a website, preceded by checking the robots.txt file and server headers
- Essential steps:
    - Identify the target population
    - Obtain the URLs of the units to collect data from
    - Make a request to these URLs to get the associated HTML-page (incl. check robots.txt)
    - [use specific locators to find the part of interest in the HTML-code]
    - Save selected data

Population perspective

- Depending on the topic of interest, various numbers of websites need to be scraped
- Not all units of interest may actually have a website

Table 1. Web scraping examples by population size

Population size	Examples
P1: One website	Satellite data Search engine results
P2: Selected websites (Purposing sampling)	Online Job Advertisements Real estate prices Price statistics
P3: All websites	Enterprise characteristics Innovative company detection

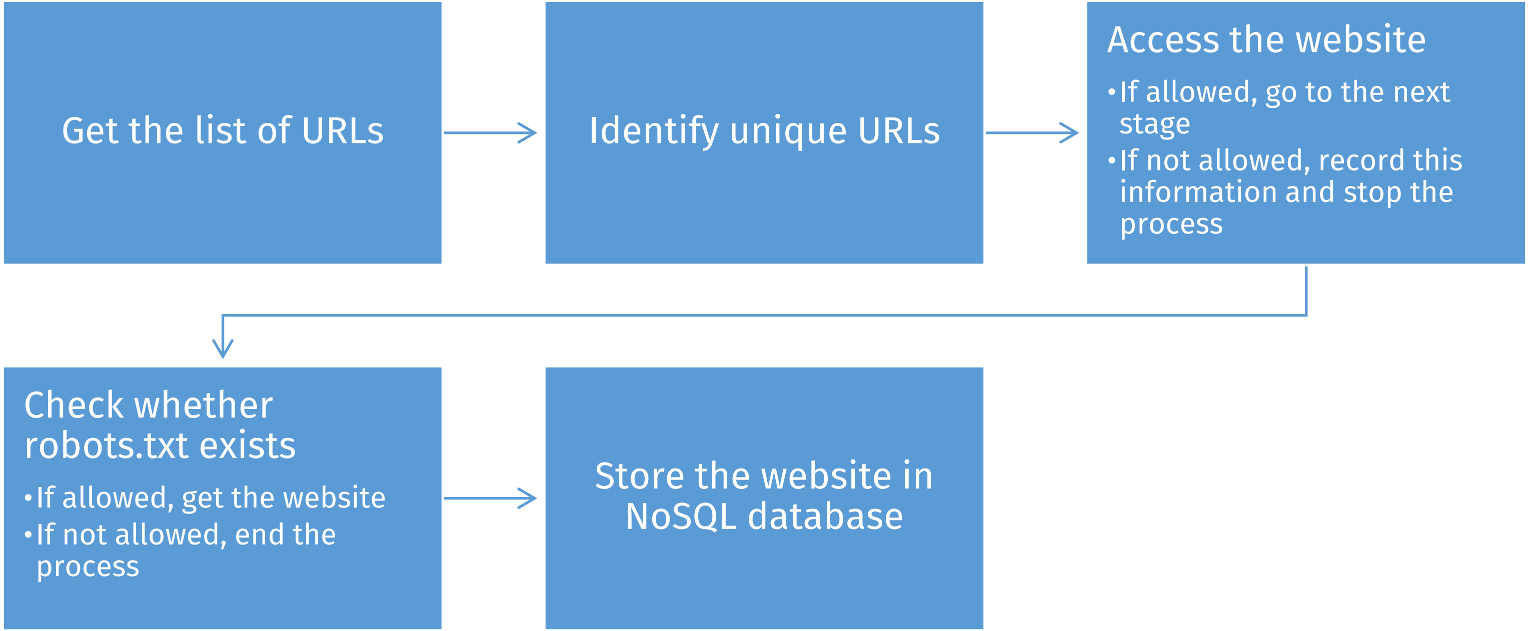
Source: own elaboration.

URLs perspective

- For the units identified, the URL of their website need to be identified
- It can be done by:
  - Find and re-use existing URL databases
  - Buy from third-party companies (external)
  - Find URLs with URL-finding approach
- Not all URLs can be accessible

Web scraping process used in the case study

Figure 1. Web scraping process used in the case study



Get the list of URLs

Identify unique URLs

Access the website

- If allowed, go to the next stage
- If not allowed, record this information and stop the process

Check whether robots.txt exists

- If allowed, get the website
- If not allowed, end the process

Store the website in NoSQL database

Source: Own elaboration.

Data collection

- The URL data for the Dutch study were obtained from the Dutch company DataProvider which were subsequently linked to the corresponding businesses in the Business Register of Statistics Netherlands at the most detailed level possible.

- Database for Polish enterprises was used from Bureau van Dijk database Orbis.

Table 2. Results of the case study

Specification	Number of websites
Population size	503,700 (100%)
Unique domain names	459,700 (91%)
Accepted connections	340,700 (74%)

Source: own elaboration.

Issues identified during case study [1/2]

Table 3. Problems identified during the process of URL sampling and web scraping

No.	Issue	Methods of mitigating
1	Incomplete URL list	Use URL search to find additional URLs
2	List of URLs has non-updated data	Use URL search script to verify if URLs have changed
3	Not up-to-date information on websites	Regularly scrape websites
4	Website is blocking robots	Try to use an alternative approach, i.e. use different web browser engine, to scrape data and inform website owner of the issue
5	Robots.txt rejection	Inform website owner of intention to scrape the data (scrape anyway)
6	Temporary unavailability	Attempt to scrape website at another time/date

Source: own elaboration.

Issues identified during case study [1/2]

Table 3. Problems identified during the process of URL sampling and web scraping

No.	Issue	Methods of mitigating
7	No time stamps	Regularly scrape website and monitor changes by comparing stored data in NoSQL database
8	Duplicates of websites	Include de-duplication mechanisms, include URL-forward checks
9	Only partial information obtained	Check if website is still active and, if that's the case, check script to extract more data
10	The quality of the link between an enterprise and the URL	Check whether the website refers to the enterprise in the population by verifying that company details, like name or address exists in the content of the website
11	Information on enterprises without a website (if relevant)	Check whether there are other sources of information available, such as a survey, or contact a small sample to obtain an indication of the number of enterprises and type(s) of data missing

Source: own elaboration.

Web scraping main findings

- When we don't know how many businesses have a website, we can estimate this number using web-scraped data.
- However, because this type of data is often biased, our estimate may not be accurate.
- This affects the findings as the coverage of some classes of enterprises could be over/underestimated.
- Therefore, a ‘survey’, understood as a questionnaire with questions to be answered, based on web data may provide data aggregates that do not accurately represent the intended target population.

Conclusions and final remarks

- Finding complete set of URLs is very unlikely
- Methodology is needed to deal with data collection and quality issues as good as possible
  - Integrating traditional surveys with web surveys
  - Non-probability sampling
- Currently our results are used as a methodological input for webscraping for Web Intelligence Hub/Network use case:
  - Online Based Enterprise Characteristics
- Privacy issues and GDPR are a key issue when doing massive web scraping

Thank you for your attention!

Jacek Maślankowski, j.maslankowski@stat.gov.pl
Piet Daas, pjh.daas@cbs.nl