From web to data:
Selective scraping and WIN.The hackathon

Olav ten Bosch, Statistics Netherlands
o.tenbosch@cbs.nl

1. INTRODUCTION

Web data is becoming increasingly important for official statistics. In combination with traditional inputs web data may add up-to-date information to the existing data portfolio, it may improve or speed-up statistical processes or it may creating new opportunities for (experimental) indicators. There are successful examples in price statistics on web shops, enterprise statistics on business websites and social statistics observing social media. However, there are also challenges: web data may be volatile, may contain multiple types of biases and quality may depend on many factors, such as the business value of the data for the content provider.

In the European ESSnet project Trusted Smart Statistics – Web Intelligence Network (TSS-WIN) [1] multiple ways of using web data in official statistics have been explored. Apart from major use cases such as online job advertisements (OJAs), online business enterprise characteristics (OBECs) a work package was dedicated to exploring other use cases, such as characteristics of the real estate market, construction activities, online prices of certain product categories, indices in tourism statistics. Another major use case is business register quality enhancement. They all had their successes and challenges such as volatile inputs, deduplication challenges, mapping web data onto statistical concepts and operational challenges, but also helped further shape the relatively new thinking about *selective scraping versus bulk scraping*, first explained in [2].

2. THE CONCEPT OF SELECTIVE SCRAPING

In many cases the use of web data starts with collecting initial samples of data from one or more websites that seem valuable for the statistical use case at hand. The websites are studied, some data is collected and analysed and, if found to be valuable, data collection is expanded to cover the web domains needed for the quality desired. This may involve gathering raw data from various online sources and subsequently linking it to the relevant statistical context. This can be a labour-intensive process, especially as new data sources emerge and the volume of data increases. We call this method bulk-scraping.

Statistical[1] or selective scraping, on the other hand, leverages the existing knowledge base of statistical offices. By querying the web with specific identifiers, names, categories, or statistical definitions, already contained in the registers

[1] In [2] the term “statistical scraping” is used. However “selective scraping” seems to become more popular and also adheres to the well-known “selective editing” in editing in official statistics. Hence, in the sequel of this abstract we will use the term selective scraping.

maintained by the statistical organisation, it ensures that the collected data can be directly linked to the statistical context. This approach is akin to conducting an automated survey on the vast expanse of the web, where the results, though potentially messy, can be connected to the relevant statistical units. By applying such a targeted approach to data collection, data volumes are limited, data in registers are maximally used and statistical agencies can reduce the risk of representation errors and calculate accuracy and reliability of their findings. Figure 1 sketches the principle.

It is important to note here that bulk scraping certainly has its value in certain statistical use cases. Even stronger, in cases where the statistical population yet has to be discovered it may be the only option. However, in other cases a selective scraping methodology may complement or in some cases replace bulk scraping methods.

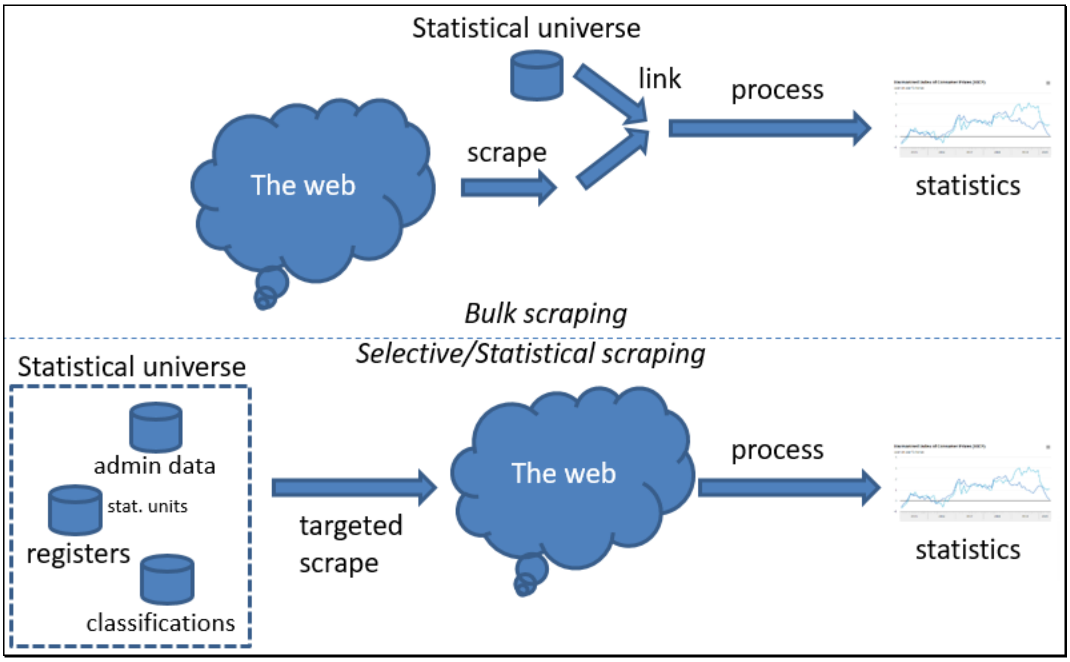

Figure 1. Selective scraping versus bulk scraping

3. BUSINESS REGISTER ENHANCEMENTS FROM WEB DATA

One notably example of selective scraping is improving the business register with web data, one of the use cases studied in the WIN project. Statistical Business Registers (SBRs) are indispensable assets for National Statistical Institutes (NSIs). They provide detailed information about enterprises, serving as the foundation for numerous statistical analyses and surveys.

Online data offers a wealth of insights into the activities of enterprises. Websites, media advertisements, product listings, customer interactions, Wikipedia articles, job postings can all be usual information to better classify the statistical unit or to check or complete the information available in the SBR.

Using selective scraping, the first step in this process is to identify the URLs associated with each enterprise. This can be challenging as URLs are often not readily available or may be unreliable. To address this, NSIs can utilize search engines using information from the SBR and select the most relevant matches based on factors such as the enterprise's name, contact information, and tax identification number.

Once URLs are identified, the next phase involves extracting statistical variables from the associated web contents. This typically involves scraping text and applying natural language processing (NLP) and machine learning techniques to interpret the content. Common use cases include deriving economic activity (NACE), ecommerce, social media use, enterprise relationships, sustainability practices, and / or job vacancies. Of course, this is a simplification, in reality the relationships between legal units in the SBR and websites or other web data may be more complex. A single enterprise may operate multiple websites for different customer segments, while a website may be used by multiple legal units within an enterprise group. Nevertheless, generally speaking selected scraping for enhancing business registers has proven to be a valuable concept.

4. Other examples

Other domains where the selective scraping concept has been applied or can be applied are 1) price statistics (the basket approach applied to the web), 2) Tourism statistics (scraping a selected sample of hotels in certain categories) and 3) job vacancies (checking whether a selected sample of enterprises from the business register have job vacancies, how many and possibly what kind of).

5. WIN.The hackathon

Building on the selective scraping concepts explained above, the WIN project issued a hackathon on OBEC indicators in 2024. Target indicators were ecommerce and social media use, where social media use was specialized into 6 types of social media. A set of URLs, geographically spread over 4 countries (Poland, Austria, Germany and the Netherlands) was derived. A subset was manually labeled and a community of data scientists was asked to come up with the best open source software to automatically derive these indicators from websites. The results were scored against the labeled set, which was secret to the participating teams.

In this “From web to data” session the teams that were closest to the labeled results will present their results and experiences and the innovative way they came to their solution.

References

[1] WIN, the Web Intelligence Network project, https://cros.ec.europa.eu/book-page/web-intelligence-network-project-overview

[2] Ten Bosch, O., Kowarik, A., Quaresma, S., Salgado, D., van Delden, A. (2024). Statistical scraping: informed plough begets finer crops. Paper for the Q2024 conference, Estoril, Portugal. Link: https://www.researchgate.net/publication/380532685_Statistical_scraping_informed_plough_begets_finer_crops