Statistical scraping: informed plough begets finer crops

Olav ten Bosch[1], Alexander Kowarik[2], Sónia Quaresma[3], David Salgado[4], Arnout van Delden[1]

[1]Statistics Netherlands, the Netherlands

[2]Statistics Austria, Austria

[3]Statistics Portugal, Portugal

[4]Statistics Spain, Spain

Abstract

The use of web data for official statistics has been studied extensively in recent years. It is widely recognised that combining such data with traditional inputs may improve and speed-up statistics, and may open up possibilities for new indicators that couldn't otherwise be measured. There are successful examples in price statistics scraping web shops, enterprise statistics scraping business websites and social statistics using social media. However, there are also challenges: web data are volatile, rich of biases and of unknown quality, to name a few. But the biggest problem is methodological: how to link, map or cluster the web data not designed for statistics, which uses messy real-world language, into statistical units or aggregates needed for official statistics.

Traditional approaches often involve collecting raw data from various online sources exposing information related to the statistical concept of interest. Over time, new data sources are added resulting in a bulk scraping approach, where possibly voluminous data streams have to be linked to the statistical context. In contrast, statistical scraping starts at the knowledge that statistical offices already have. The web is queried with an identifier, name, category, or statistical definition, so that the result can always be linked back to the statistical context. It's like performing an automated survey on the huge web with messy but linkable results. Or like a farmer that knows from experience what fertilizer and harvesting strategy makes the finest crops. This strategy helps noticeably to cope with representation errors.

A notable example of statistical scraping is in business register enhancement, a subject explored in the European WIN project. Starting from information in the business register the web is searched to identify digital traces associated with certain statistical units. These traces are then employed to enhance administrative or statistical variables such as NACE codes. Another example can be found in price statistics. Statistical scraping in this context implies a search for well-defined products from the basket to collect high quality price observations for those products. An untapped area with regard to statistical scraping, but where it could yield valuable insights is job market statistics.

In this presentation, we sketch the concept of statistical scraping, a methodology that may complement or in some cases replace bulk scraping methods. We explore examples, consequences and its possible impact on data quality. Finally, we ponder the potential implications on ongoing and future projects that utilize web data sources for official statistics, ensuring the preservation of the high-quality standards expected for official data.

Keywords: online data, web scraping, registers, new data sources

1. Introduction

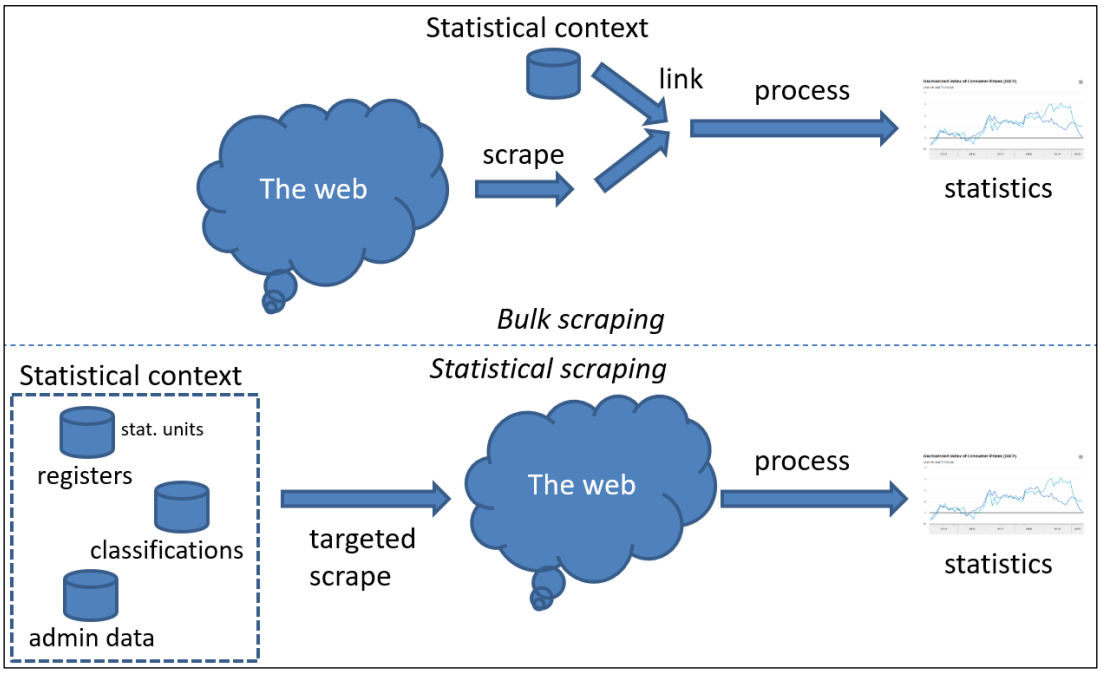
In addition to surveys and administrative sources, online data have shown to be a valuable additional input for making high quality official statistics. Over past years many projects have shown the value of the web for measuring phenomena that where difficult to observe otherwise, or could be measured more efficiently via the web (Hoekstra et al., 2012; Ahmad Yar et al., 2023, Diniz et al., 2023). The authors of this paper have been involved in quite some innovative projects and in recent discussions on European scale on the way forward using new data sources. In this paper we present our thoughts on future scraping projects, with the aim of making the addition of online data to the official statistics data menu not only efficient and manageable but also methodologically sound.

The concept we present here is the idea to design scraping projects as much as possible taking into consideration the statistical knowledge already contained in the statistical offices. Statistical offices typically maintain a multitude of meta information such as statistical registers, classifications, code lists and data derived from administrative sources. They use it as the statistical context in which data, varying from surveys to big data, are processed into meaningful statistics (Reister, 2023). As opposed to bulk scraping where large volumes of data are taken from online sources and linked to the statistical context, "statistical scraping"[1] starts with the domain knowledge on the statistical domain to perform a targeted scrape. This can be searching for a unit in a statistical register, or scraping (parts of) websites which have been selected after a landscaping exercise (Six and Kowarik, 2023) based on statistical knowledge, or scraping selected sources that cover a specific category in a classification.

Figure 1 sketches the concept on a high level. In bulk scraping web data is scraped and linked to the statistical context at hand a-posteriori. Apart from practical disadvantages on retrieving and managing larger volumes of web data, one of the main methodological disadvantages here is that it invites representation errors if the target population is partly or entirely dependent on the scraped information. In order to overcome these disadvantages, we propose *statistical scraping*. Statistical scraping uses a-priori information, e.g., the knowledge contained in registers, on unit level or higher aggregate level, in classifications or in other metadata to perform a *targeted* scrape on the web. The essential difference is that the results are inherently related to the statistical context at hand.

[1] As an alternative term 'selective scraping' has been proposed. We are neutral about the term to be used, either one describes the thinking quite well. For simplicity we will use statistical scraping throughout the rest of the paper.

Figure 1: High level view on bulk scraping versus statistical scraping



In this paper we take an explorative approach to discover the concept gradually. In section 2 we explore some of the scraping projects we faced over time and analyse their relationship with the statistical scraping concept. In section 3 we generalize the concept into a more formal definition and approach and sketch the added value to the statistical processing and methodological view. In section 4 we wrap up the ideas presented.

2. Discovering the concept by example

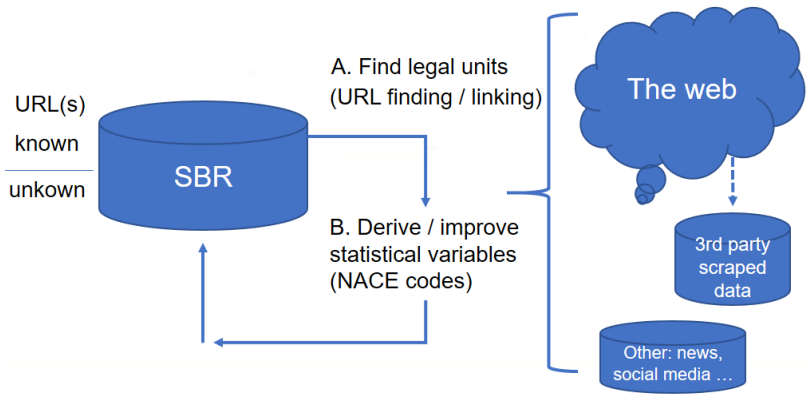
2.1 Business register improvement

Statistical business registers (SBRs), databases of statistical units such as enterprises, are important assets for National Statistical Institutes (NSIs). These registers contain detailed information about enterprises with variables such as administrative details, location, size, economic activity, ownership information and information on the legal composition and possibly relationships with other units. The SBR defines the target population for many statistics and serves as frame for statistical surveys. In the international Web Intelligence Network project (WIN consortium, 2024), one use case is concerned with business register quality enhancements from online data.

Online data comprises many forms. In fact, every digital trace left by a company, such as websites, media advertisements, product listings, customer care interactions, Wikipedia data, and job postings, may offer valuable insights usable for business register enhancements.

These data sources are explored using search, link and scrape techniques (ten Bosch et al., 2023). Although all digital traces are potentially interesting, enterprise website(s) have special focus. Unfortunately, the URL(s) of the enterprise is in many cases not known. If missing, or not reliable enough, they must be found in the so-called URL finding phase (A). In the next phase (B), statistical variables can be derived or improved from the data found. Figure 2 shows the approach visually.

Figure 2: High level view on statistical business register (SBR) improvements via search, scrape and deriving indicators using machine learning techniques.



One way to do URL finding (A) is to use one or more (paid) search engines. One or more search queries are executed containing the name of the unit, optionally supplemented with contact information and / or tax identification number. The results are interpreted and weighted using a machine learning model trained on the specifics of the respective engine to select the best match(es). The set of legal units in the business register with known URLs may serve as a training set (van Delden et al., 2019). Two other ways of URL finding are the use of web data collected by third parties or using the domain registry of the respective country.

The second phase (B) is to derive statistical variables from online data. This is typically done by scraping text from the URLs found, keeping the relationship with the unit in the SBR. Deriving economic activity (NACE), which involves interpretation of raw texts with natural language processing (NLP) and machine learning, is a common use-case (Mangat et al., 2023). Other examples are ecommerce, social media use, degree of sustainability or job vacancies. Yet another use case is to discover or verify administrative information of SBR units, such as email addresses and names of managers of businesses. In both phases, A and B, special attention should be paid to many-to-many relationships between legal units and websites. So a legal unit or enterprise might operate several websites, e.g., for different client

segments, and a website can be collectively used by several legal units, e.g., within an enterprise group.

This example of business register enhancement from online data is a prototypical example of statistical scraping. In the URL finding phase information on unit level to perform a targeted scrape mostly using search engines. In phase B scraping is performed targeted on the specific website(s) which are found to belong to the statistical unit in the business register. All in all, the whole process of online data exploitation is driven from the statistical context maintained within the statistical office.

2.2 Prices statistics

The use of web data in price statistics has been described in many publications, from the early experiments (Cavallo, 2009) to mature applications in official statistics (Griffioen and ten Bosch, 2016) to specific national approaches (Manik and Albarda 2015; Oancea and Necula 2019). The early approaches were typical bulk scraping applications retrieving prices in large which then were aggregated into an price index, with the obvious disadvantage of possibly not being representative for the actual budget patterns that the CPI is supposed to measure. These days most of the approaches scrape a subset of products relating to certain categories in the statistical classification for individual consumption, COICOP. The sites to be scraped are determined from a landscaping exercise in which representative websites are selected.

Another form of using online data for price statistics, which was called robot-assisted data collection, has been described by ten Bosch and Windmeijer (2014). This approach is the reincarnation of the traditional basket-based price collection but now applied to the web. The basket contains well-defined products with their respective weights. For each product a number of measuring spots on the web are defined and regularly visited. This process is supported by a tool which is now in production for over 10 years. Later, a more advanced variant has been designed where the products are discovered by site-specific or generic search engines.

Both the landscaping-based price scraping as well robot-assisted collection are examples of statistical scraping. In the latter, web content is collected on the lowest product classification level. Landscaping based COICOP scraping is an example of statistical scraping on classification-level. Both are driven from a statistical context.

2.3 Other examples

Since nowadays most of the hotel and other travel arrangements are made online, tourism statistics is a field where online data plays an increasingly important role. In addition to bulk scraping methods where data from International tourism accommodation websites are collected and aggregated, an alternative method has been used, based on hotel star ratings. In a first step for each star 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. This survey-like approach to online data is an excellent example of statistical scraping in the tourism domain.

Coffey et al. (2024) contains some more examples of using web data as auxiliary data in their “survey-plus environment”. One example is the use of web data starting from a register of schools to draw a sample of teachers (Avenilla, 2022), which is a targeted scrape from a statistical context and as such another good example of statistical scraping.

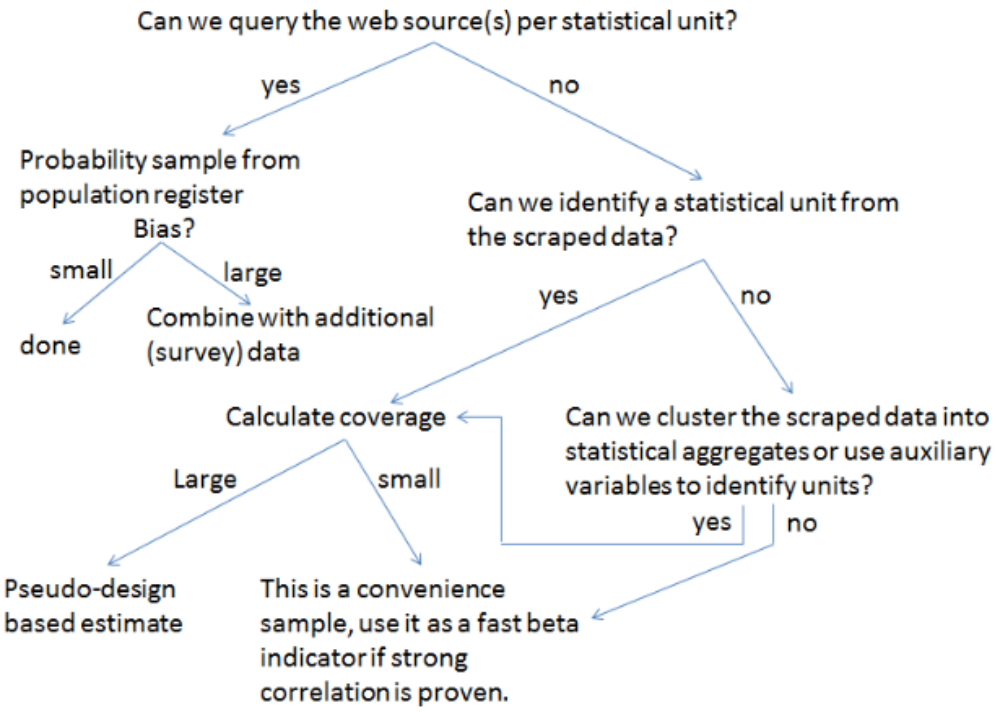
A last example may be found in job vacancy statistics on online data. Up to now, most approaches (Cedefop, 2019; Romanko and O’Mahony, 2022) focus on a bulk scraping strategy. However, it has been shown that the relationship from the online job advertisement with the enterprise is difficult to derive. A statistical scraping approach would start at the business register, visiting the website(s) registered with the unit at low frequency to derive the number and types of job vacancies open at time of inspection. The scraper could ultimately arrive at the same job portal as with bulk scraping, but now with the valuable connection with the statistical unit at hand.

3. A more formal view on the concept

3.1 Definition

In (ten Bosch et al., 2018) a first version of a generic workflow was proposed that could be applied to any web source in the scope of official statistics. This workflow is repeated in Figure 3.

Figure 3: Generic workflow for using web data in the scope of official statistics.



For a full explanation of the individual steps we refer to the original publication (pp. 9-11). For the definition of statistical scraping, we focus on the first step: the question whether one can query the web source(s) per statistical unit. Statistical scraping generalizes this step. It is not only applicable on statistical unit level, essential to the approach is the connection with the statistical context from which we make statistics. Hence, we come to the following 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.*

3.2 Methodological consequences

With scraping approaches one often has to cope with linking errors (the link from the scraped content with the target population fails), selection errors (not for all units in the target population it is possible to retrieve data, which may lead to unknown biases), and frame errors (the statistical registers contain errors too). Statistical scraping as opposed to bulk scraping helps to better cope with the first two types of errors, since the relationship with the statistical context, at unit level or (sub)aggregate level is known.

For instance, with business registered improvement (see section 2.1) one typically starts with finding the URL(s) of the enterprises in the SBR, using a variety of data sources. The estimated linkage probability is an example of the relationship with the statistical context. In the next step one can then decide to scrape all URLs of a certain minimum linkage

probability, which is an example of starting from a-priori information in this statistical domain. Bulk scraping typically starts with one of the possible URL sources leading to a higher selectivity if one *only* relies on that source. The linkage probability of part of the units obtained by bulk scraping may also be small especially when contact information is missing on the scraped websites.

If statistical scraping can be applied on the unit level, which depends obviously on the query possibilities of the statistical object at hand and the web data source(s) to be queried, well-defined and proven survey methodology quality indicators can be calculated. Examples of such quality indicators are the linkage probability between URL and business (legal unit), and under coverage: the proportion of units of the target population not covered by the web source.

All in all, statistical scraping provides an approach to cope with different representation errors by making use of the available information regarding the target population under analysis.

3.3 Other consequences

Statistical scraping can be seen as retrieving exactly those data that is needed for the statistical task at hand. Hence, it may involve smaller data streams, with less pressure on hardware and bandwidth, and lower scraping frequencies, which makes it more manageable. On the other hand, statistical scrapers need to be aware of the statistical context they operate from, which may need a more secure scraping setup where possible sensitive input data is to be handled with care.

When a national statistical institute is scraping web data for statistical purposes, the GDPR principle of data minimization is relevant. Since the data being scraped include personal data, e.g. as contact information on enterprise websites, the NSIs must ensure that only the minimum amount of personal data necessary for their statistical purposes is collected.

If search engines are used, identifying information contained in a query can be identified in web server logs (search engine leakage). This risk can be reduced by carefully designing the queries, spreading them across different search engines, or entering into a non-disclosure agreement with a search provider. Another intriguing thought is that the official statistics community would create its own safe search engine or use one in partnership with EU initiatives (Granitzer et al., 2023). Either approach will do, and all in all, we think that search engine leakage is to be taken serious but manageable.

From a generic standpoint regarding the production of official statistics within the new data ecosystem, statistical scraping helps underline the difference between data and statistics

(Reister, 2023) through the integration of web data with existing information in the different official statistical domains resonating the importance of a methodologically sound production process and the assurance of quality of the final statistical products.

4. Wrap-up

Online data have shown to be a valuable new data source for the production of official statistics, either as primary or auxiliary data source. In addition, to bulk scraping approaches, where potentially voluminous data is collected from various data portals, a statistical scraping approach is presented. In statistical scraping the relationship with the statistical context is contained. This approach uses the knowledge contained in registers, on unit level or higher aggregate level, in classifications or in other metadata to perform a targeted scrape on the web, which makes the results inherently related to the statistical context at hand.

Examples can be found in business register enhancement, where the information in the statistical business register is used as a starting point to consult digital traces of enterprises to improve the register, in price-statistics, where a landscaping or basket approach can be applied to find and collect exactly what is needed, in tourism, where a probability sample of accommodations is scraped with low frequency, and in drawing a sample of school teachers from a collection of school websites searched from register data. An example that has yet to be explored is to use statistical scraping starting from the business register for job market statistics.

Based on a definition of statistical scraping in this paper, some of the methodological consequences have been explored, such as the possibility to calculate survey methodology quality indicators and organisational and privacy consequences, such as the careful treatment of possible sensitive scraping input data. Concluding, we think that for future scraping projects it is wise to, in addition to bulk scraping methods, consider a statistical scraping approach where, depending on the statistical knowledge contained in the statistical office and the querying possibilities of the data sources at hand, selected samples with high statistical value are collected. There are still many choices to make and experiences to gain with this approach, but we hope that the reasoning in this paper helps in our collective future innovation exercises.

Acknowledgment

The views expressed in this article are those of the authors and do not necessarily reflect the policies of their institutes. We thank all colleagues in our institutes involved in web scraping projects. We also thank our colleagues in the European TF Trusted Smart Statistics (TF-TSS) for fruitful discussions, which helped formulating the ideas presented here.

References

- Ahmad Yar, A. W., and Bircan, T. (2023). Big data for official migration statistics: Evidence from 29 national statistical institutions. *Big Data & Society, 10(2)*. https://doi.org/10.1177/20539517231210244
- Cedefop (2019). Online job vacancies and skills analysis – A Cedefop pan-European approach. *EU Publications Office*. https://data.europa.eu/doi/10.2801/097022
- WIN consortium. (2024). *ESSnet Web Intelligence Network*. https://cros.ec.europa.eu/landing-page/web-intelligence-network
- Avenilla, L. (2022). NTPS Web Scraping. In *Presentation at the Federal Committee on Statistical Methods Conference, Washington, D.C., October 27*. https://www.fcsm.gov/assets/files/docs/2022-conference-docs/H1.4_Avenilla.pdf
- Cavallo, A. (2009). *Scraped Data and Sticky Prices: Frequency, Hazards, and Synchronization*. Harvard University.
- Coffey, S. M., Damineni, J., Eltinge, J., Mathur, A., Varela, K., and Zotti, A. (2024). Some Open Questions on Multiple-Source Extensions of Adaptive-Survey Design Concepts and Methods. *JOS*, *40*(1), 16–37. https://doi.org/10.1177/0282423x241235270
- Granitzer, M., Voigt, S., Fathima, N. A., Golasowski, M., Guetl, C., Hecking, T., and Zerhoudi, S. (2023). Impact and development of an Open Web Index for open web search. *Journal of the Association for Information Science and Technology*. See also https://openwebsearch.eu/
- Griffioen, R., and ten Bosch, O. (2016). On the use of Internet data for the Dutch CPI. *UNECE - Meeting of the Group of Experts on Consumer Price Indices, Geneva, 2016*
- Hoekstra, R., ten Bosch, O., and Harteveld, F. (2012). Automated data collection from web sources for official statistics: First experiences. *Statistical Journal of the IAOS*, *28*(3, 4), 99–111. https://doi.org/10.3233/SJI-2012-0750
- Mangat, M., Gussenbauer, J., and Kowarik, A. (2023). Using Webdata to derive the Economic Activity of Enterprises. *UNECE Machine Learning for Official Statistics Workshop '23, Geneva*
- Manik, D. P. and Albarda (2015). A strategy to create daily Consumer Price Index by using big data in Statistics Indonesia. *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, 1–5.
- Oancea, B. and Necula, M. (2019). Web scraping techniques for price statistics–the Romanian experience. *Statistical Journal of the IAOS*, *35*(4), 657–667.
- Reister, M. (2023). Assuring quality in the new data ecosystem: Mind the gap between data and statistics! *Statistical Journal of the IAOS, 39(2)*, 421-430.
- Romanko, O. and O'Mahony, M. (2022). The use of online job sites for measuring skills and labour market trends: A review. *Economic Statistics Centre of Excellence (ESCoE) Technical Reports (ESCOE-TR-19)*. http://escoe-website.s3.amazonaws.com/wp-content/uploads/2022/05/30133155/TR-19.pdf
- Six, M. and Kowarik, A. (2023). Issue 10 - Quality aspects of web scraped data - Focus on landscaping and selection of sources. Web Intelligence Network Blog, https://cros.ec.europa.eu/book-page/issue-10-quality-aspects-web-scraped-data-focus-landscaping-and-selection-sources
- ten Bosch, O., van Delden, A., and de Wolf, N. (2023). Business register improvements: a balance between search, scrape and 3rd party web data. *NTTS 2023*.
- ten Bosch, O., and Windmeijer, D. (2014). On the use of internet robots for official statistics. *UNECE Meeting on the Management of Statistical Information Systems (MSIS), 2014*.
- ten Bosch, O., Windmeijer, D., van Delden, A., and van den Heuvel, G. (2018). Web scraping meets survey design: combining forces. *BigSurv18, Barcelona, Spain*. https://www.researchgate.net/publication/327385487_Web_scraping_meets_survey_design_combining_forces
- Van Delden, A., Windmeijer, D., and Ten Bosch, O. (2019). Finding enterprise websites. *European Establishment Statistics Workshop. Bilbao, Spain*. Available at: https://www.researchgate.net/profile/Arnout_Delden/publication/336995371_Finding_enterprise_websites/links/5dbebf29299bf1a47b0f5669/Finding-enterprise-websites.pdf