Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 49–64 https://doi.org/10.59139/ws.2023.12.3
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 13.09.2022, poprawiony/revised: 28.08.2023, zaakceptowany/accepted: 20.09.2023

Current challenges and possible big data solutions for the use of web data as a source for official statistics[1]

Piet Daas,[a] Jacek Maślankowski[b]

Abstract. Web scraping has become popular in scientific research, especially in statistics. Preparing an appropriate IT environment for web scraping is currently not difficult and can be done relatively quickly. Extracting data in this way requires only basic IT skills. This has resulted in the increased use of this type of data, widely referred to as big data, in official statistics. Over the past decade, much work was done in this area both on the national level within the national statistical institutes, and on the international one by Eurostat. The aim of this paper is to present and discuss current problems related to accessing, extracting, and using information from websites, along with the suggested potential solutions.

For the sake of the analysis, a case study featuring large-scale web scraping performed in 2022 by means of big data tools is presented in the paper. The results of the case study, conducted on a total population of approximately 503,700 websites, demonstrate that it is not possible to provide reliable data on the basis of such a large sample, as typically up to 20% of the websites might not be accessible at the time of the survey. What is more, it is not possible to know the exact number of active websites in particular countries, due to the dynamic nature of the Internet, which causes websites to continuously change.

Keywords: big data, web data, websites, web scraping

JEL: C55, L86, M21

Współczesne wyzwania i możliwości w zakresie stosowania narzędzi big data do uzyskania danych webowych jako źródła dla statystyki publicznej

[1] Artykuł został zaprezentowany w postaci referatu na konferencji *Metodologia Badań Statystycznych MET2023*, która odbyła się w dniach 3–5 lipca 2023 r. w Warszawie. / The article was presented in the form of a lecture at the *MET2023 Conference on Methodology of Statistical Research*, held on 3rd–5th July 2023 in Warsaw.

[a] Eindhoven University of Technology, Department of Mathematics and Computer Science, the Netherlands. ORCID: https://orcid.org/0000-0002-1541-0315. E-mail: p.j.h.daas@tue.nl.

[b] Uniwersytet Gdański, Wydział Zarządzania; Urząd Statystyczny w Gdańsku, Ośrodek Inżynierii Danych, Polska. / University of Gdańsk, Faculty of Management; Statistical Office in Gdańsk, Centre for Data Engineering, Poland.
ORCID: https://orcid.org/0000-0003-0357-2736. Autor korespondencyjny / Corresponding author, e-mail: jacek.maslankowski@ug.edu.pl.

Streszczenie. Web scraping jest coraz popularniejszy w badaniach naukowych, zwłaszcza w dziedzinie statystyki. Przygotowanie środowiska do scrapowania danych nie przysparza obecnie trudności i może być wykonane relatywnie szybko, a uzyskiwanie informacji w ten sposób wymaga jedynie podstawowych umiejętności cyfrowych. Dzięki temu statystyka publiczna w coraz większym stopniu korzysta z dużych wolumenów danych, czyli big data. W drugiej dekadzie XXI w. zarówno krajowe urzędy statystyczne, jak i Eurostat włożyły dużo pracy w doskonalenie narzędzi big data. Nadal istnieją jednak trudności związane z dostępnością, ekstrakcją i wykorzystywaniem informacji pobranych ze stron internetowych. Tym problemom oraz potencjalnym sposobom ich rozwiązania został poświęcony niniejszy artykuł.

Omówiono studium przypadku masowego web scrapingu wykonanego w 2022 r. za pomocą narzędzi big data na próbie 503 700 stron internetowych. Z analizy wynika, że dostarczenie wiarygodnych danych na podstawie tak dużej próby jest niemożliwe, ponieważ w czasie badania zwykle do 20% stron internetowych może być niedostępnych. Co więcej, dokładna liczba aktywnych stron internetowych w poszczególnych krajach nie jest znana ze względu na dynamiczny charakter Internetu, skutkujący ciągłymi zmianami stron internetowych.

Słowa kluczowe: big data, dane webowe, strony internetowe, web scraping

1. Introduction

The use of web-scraped data for the production of official statistics encounters numerous methodological challenges. When the number of businesses maintaining websites is unknown, we can estimate it using web-scraped data. However, because this type of data is often biased, our estimate may not be accurate, i.e. some classes of enterprises could be over- or underestimated. Therefore a survey, understood as a questionnaire with questions to be answered, which is based on web data, may provide data aggregates that do not accurately represent the intended target population.

Web scraping refers to the process of using software for automatic extraction of data from websites (Khder, 2021). In this paper, we understand web scraping as a method to get the source of a website, i.e. the source file (mostly HTML), preceded by checking the robots.txt file and server headers. Web-scraped data are extracted from the website to get useful information.

The aim of this paper is to present and discuss current problems related to accessing, extracting, and using information from websites, along with the suggested potential solutions. The secondary aim is to provide an overview of methods and cases that can be used and replicated to extract statistical data from websites. This article is the result of the authors' long experience in working with this type of data.

The essential research question in this study is whether it is possible to collect internet data suitable for the production of official statistics using massive web scraping techniques. Along with the literature review, the research methods used in the paper comprised the authors' case studies of enterprise websites and case studies conducted at the European level, all focused on producing official statistics. In this

paper we demonstrate the results of a case study involving massive web scraping, performed in 2022 on a total population of 503,700 Polish websites. The results showed that such a large sample could not yield fully reliable data, as usually up to 20% of the studied websites are not accessible at the time of the survey. These findings are in line with other studies in this area, for example Oancea and Necula (2019) or Daas and van der Doef (2020). Web-scraped data has been regarded as a data source for official statistics for more than 10 years (Daas et al., 2015). Nowadays, web-scraping is a fundamental requirement during data scientist training (Dogucu & Çetinkaya-Rundel, 2020). This technique has not only been adopted in statistical research, but also in scientific papers or marketing reports, which includes, for example, marketing scholars using Application Programming Interface (APIs) to collect data from the Internet. In marketing research, the number of papers using online data increased from 1% in 2001 to 15% in 2020 (Boegershausen et al., 2022). Researchers relatively often collect price data from the web, for example, to calculate the value of the real estate market (Antonov & Laktionova, 2020), to produce price indices of real estate (Pegueroles et al., 2021) and used cars (Nasiboglu & Akdogan, 2020), to compile an experimental consumer price index (Oancea & Necula, 2019), or to produce consumer electronic products (goods) and airfares (services) price indices in order to improve the Harmonised Index of Consumer Prices (Polidoro et al., 2015). Another well-known example is the Billion Prices Project at the Massachusetts Institute of Technology (MIT), which scrapes massive amounts of prices from the web to produce daily online price indices for the USA and several other countries (Cavallo & Rigobon, 2016). Financial and other types of information can easily be extracted from web data with a variety of supporting packages, which provide basic tools used for pre-processing web data (Krotov & Tennyson, 2018).

More advanced examples of research into web-scraped data include the use of text mining and Natural Language Processing techniques (NLP) to study local policies (Anglin, 2019) or to identify particular types of enterprises (Daas & van der Doef, 2020). In the latter case, machine learning methods are used to find words that correlate with a particular type of enterprise, e.g. innovative companies. NLP has been applied, for example, in analyses of the labour market by studying online job advertisements. Usually, a large portion of information, such as skills required for a certain job advertised online, is extracted from unstructured descriptive texts (Schedlbauer et al., 2021). Online job advertisement data can also provide input for different indicators on labour market statistics, such as the Labour Market Concentration index (Ascheri et al., 2022).

Web scraping for official statistics has been particularly well studied in a number of ESSnet projects: *Big Data I* (European Commission, n.d. a) and *Big Data II* (European Commission, n.d. b). They yielded some experimental statistics using

web data on online job vacancies, enterprise characteristics, and innovative tourism statistics (European Commission, n.d. c). The varied and complex use of web data in official statistics necessitated creating a web-scraping policy, which was formulated at the European or the NSI level (European Commission, n.d. d). It features the principles of web scraping according to good practice, such as delaying accessing pages on the same domain or adding idle time between requests (Office for National Statistics, n.d.). Currently, most of the work regarding the use of web data at the European level is done in the Web Intelligence Hub (WIH; Wirthmann & Reis, 2021) project, which is supported by an international community of statisticians within the WIN. The latter is a centralised repository offering services used to scrape data, store them in a repository, and provide data processing and analysis tools. One of the goals of the WIN is to improve knowledge and strengthen web intelligence competencies of statisticians in the use of the WIH services across the ESS and beyond (European Commission, n.d. e).

2. Research method

There are different approaches to defining statistical populations while using web data. These can be divided into three groups, as presented in Table 1.

Table 1. Web scraping examples by population size

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

Source: authors' work.

Often a single website (P1) is scraped (single-site web scraping) – for example, search engine results are collected for an analysis. In such a case, the web scraping software is collecting data from an individual website represented by one URL, i.e. a website address.

Collecting data from a set of websites selected by researchers (P2) is called purposive sampling (Palys, 2008). It involves the selection of a sample on the basis of the researcher's judgement as to which subjects fit the criteria of the study best (Purposive sampling, n.d.). Price statistics, for example, are based on a number of specific e-commerce portals. A collection of job advertisements compiled from

several websites is another example of P2. Since the sample of the scraped websites is selected before the research is performed, it may not be statistically representative of the target population the researcher had in mind, which is a potential source of bias that might seriously affect the outcome.

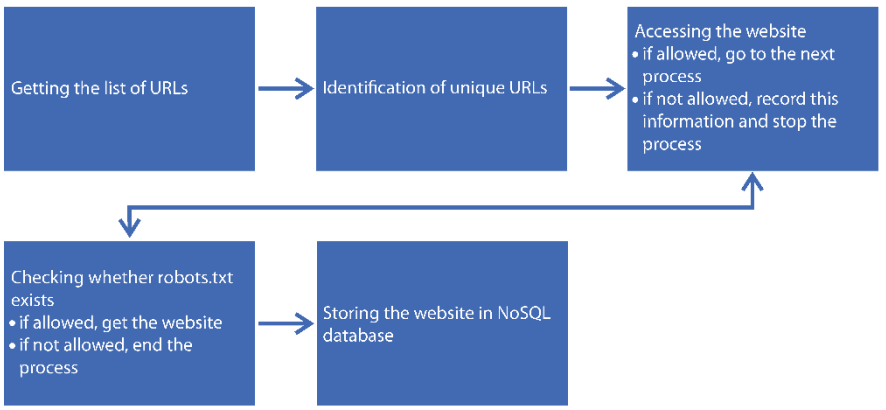
Collecting data on the entire population (P3) is the most technically-challenging approach, for three reasons. Firstly, a publicly available database of all active websites is not available in every country. There are domain registration authorities headed by the Internet Assigned Numbers Authority (IANA), but their databases are not publicly available. Secondly, creating a complete database by combining all URLs from various sources is a good starting point, but most probably, some websites will still be missing. Hence additional search and crawl process is required. Crawling refers to the process of finding additional URLs on websites already accessed. This step can also be used to check the quality of the link between the statistical units, e.g. enterprises, and websites found, and may also provide information on the units without a website. Thirdly, the composition of active websites is very dynamic; data yielded may already be outdated by the time the process is finished.

For all these reasons, obtaining an almost full overview of all websites in a country requires considerable effort. What is more, the total population of websites is often not exactly known. The size of the population is therefore often just estimated, and usually only popular websites are selected, potentially introducing biases. This is because some less popular websites, yet important from the point of view of the study, are skipped. How this can be avoided, for instance, was described in Daas and van der Doef (2020).

3. Case study on web scraping of the selected collection of websites

This section presents the case study of collecting data on a selection of enterprises. The study is based on the data scraped from the websites of Polish enterprises (URLs gathered from the Orbis database), but we will also be referring to the Dutch experience. The 'Polish enterprise' is understood as a company present in the Orbis database, located in Poland. The URL data for the Dutch study were obtained from DataProvider (a Dutch company). These data were subsequently linked to the corresponding businesses in the Business Register of Statistics Netherlands at the most detailed level possible. The linking procedure compared the Chamber of Commerce number and address from the website with those in the Business Register; see Daas and van der Doef (2020) for more details. The process used for web scraping in this case study is presented in Figure.

Figure. Web-scraping process used in the case study



Source: authors' work.

First, we decided to create a list of URLs based on the Orbis entities database managed by Bureau van Dijk. It is a commercial data provider service which maintains the database. Orbis is the repository for entity data. It consists of information on nearly 450 million companies and entities around the world (Orbis, n.d.). We took from this database all URLs linked with companies in Poland. As presented in Table 2, the collection of URLs for Poland was comparatively large (503,700 enterprises) and linked to enterprise business IDs. The linking process was based on the attributes such as an ID, name, address, etc., included in the Orbis database. Next, the set of URLs was limited to those with unique domain names, to prevent the same domain from being scraped multiple times. There were duplicates in the dataset, i.e. we found thousands of enterprises using the same domain name. This usually occurs when there is a consortium of enterprises with many branches, or when the local enterprise is based on franchise. Then, it was checked whether it was possible to access the website. Some of the websites were inaccessible, generally due to three reasons:

- website rejected by the server due to suspicious internet traffic (robot detected);
- the server was out of service/web domain was not active or
- a time-out of the request.

Sometimes it helps to re-visit an 'inaccessible' website at a later time. For instance, in the study performed by Daas and van der Doef (2020), inaccessible websites were re-visited at four different times. The next step was to check whether the robots.txt file, if available, allowed the website to be scraped. We found that for more than 95%

of the websites in which the robots.txt file denied scraping (at a certain level), it was still possible to collect the data, when ignoring this file for testing purposes. If this was not possible, the process was stopped and the appropriate flag was recorded in a NoSQL database. If the robots.txt file allowed access to the homepage, the homepage was stored in the database for further processing.

Even though we decided to use the URLs from the Orbis database, it is important to note that there are other ways of obtaining an URL list. One of the options is to actively search for websites with the URL retrieval software (European Commission, n.d. a, n.d. b). This is described in more detail in the subsequent paragraphs. The second option is to get various databases of URLs and merge them to have the most complete database. This involves using Wikipedia, the Whois database or publicly available business registers storing URLs, e.g. for Poland it could be the National Court Register.

Table 2. Results of the case study of web scraping of a selected collection of websites

Specification	Websites	
	number	percent
Population size	503,700	100
Unique domain names	459,700	91
Accepted connections	340,700	74

Source: authors' work.

The list of URLs taken from the Orbis database contained only websites of enterprises. The study was conducted in the first quarter of 2022. Duplicates were identified for about 44,000 enterprises, i.e. 8.7% of websites. In the unique population of URLs, nearly 26% of websites did not respond correctly. Servers failed to connect because of the three already-mentioned reasons. This shows that the URL database must be maintained and updated on a regular basis.

During the URL collecting-and-scraping process, we identified some problems and proposed solutions to them. They are presented in Table 3.

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

No.	Issue	Methods of mitigating
1	Incomplete URL list	Use URL search to find additional URLs
2	Non-updated data on the list of URLs	Use URL search script to verify if URLs have changed
3	Outdated information on websites	Regularly scrape websites
4	Website is blocking robots	Try to use an alternative approach, i.e. use a different web browser engine to scrape the data and inform the website owner of the issue

Table 3. Problems identified during the URL sampling and web scraping (cont.)

No.	Issue	Methods of mitigating
5	Robots.txt rejection	Inform website owner of the intention to scrape data (scrape anyway)
6	Temporary unavailability	Scrape the website at another time/date
7	No time stamps	Regularly scrape the website and monitor changes by comparing stored data in NoSQL database
8	Duplicates of websites	Apply de-duplication mechanisms and URL-forward checks
9	Only partial information obtained	Check if the website is still active and if yes, check the 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 the company's details, like the name or address, exist on 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: authors' work.

With regard to incomplete URL lists (Table 3 issue 1), it is possible to get more information by using additional sources. According to the report from the ESSnet Web Intelligence Network group responsible for obtaining data from the web, the number of URLs in the official statistical databases in selected ESS countries ranges between 2,000 and 20,000. These are the URLs of enterprises, governmental institutions, and other types of entities, like NGOs. It is important to note that URLs in official statistics are usually not collected on a regular basis, and this process is often supported by external software or third-party databases. In Poland, for example, there are several such databases, e.g. the CEiDG (the Central Register and Information on Economic Activity) and KRS (National Court Register), which are publicly available and used to support official statistics. However, some countries are only sourcing URLs from third-party institutions. An example of a fairly complete set of URLs is the Orbis database (European Commission, 2022a). The Orbis database is probably the most comprehensive set of URLs, however it requires paid subscription, according to the purchasing plans of Orbis (n.d.).

Another option is to create a list of URLs by using search terms such as the company's name, address, etc. in relevant search engines. A tool that uses search engines and evaluates and validates the resulting URLs is called a URL retrieval software. In this approach, URLs are directly obtained from web search engines and checked by visiting websites. Such software is using Google, Bing, Yahoo, or DuckDuckGo search engines to obtain a set of URLs based on the enterprise characteristics, e.g. the name, address, business ID. Usually, multiple URLs are

yielded by the search engine that needs to be checked to either reduce the number of possible websites or select the appropriate one (European Commission, n.d. a, n.d. b). Finding a correct URL can be especially challenging for small companies. This is even more the case if company names resemble each other or are too generic. For this reason, the URL retrieval is usually followed by a machine learning-based classification which classifies an URL as correct or incorrect along with the confidence rate, to increase the chance that the correct URL is found.

Another possibility to expand the URL list is to extract domain names from company e-mail addresses (European Commission, 2022b). This approach is particularly interesting for official statistics, as most of the surveys are conducted online, and the contact between the respondent and the NSI is via e-mail. While e-mail address domains may not directly relate to the enterprise in all the cases (e.g. some companies use gmail.com, outlook.com, etc.), they can nonetheless help to increase the total number of URLs found. The downside of any URL retrieval approach is the fact that they might be found for enterprises that actually do not have a website. It may happen when there are companies with the same name located in the same city or area. Even nowadays, this can still be the case for (some) small companies active in specific branches, such as farms. To sum up these issues, even though the use of third-party databases and URL retrieval software can support official statistics' URL databases, it takes much of work to obtain a complete and reliable set of enterprise URLs.

As mentioned before, official statistics requires a regularly updated list of URLs. Outdated URL lists (Table 3 issue 2) can create a scenario where a large part of the listed websites may not respond. A solution is to use software to check the availability of websites. However, in some cases, we experienced a situation where a website that did not respond at the beginning of the data collection period was active the following week. During our experiments, we established that when a website that does not respond after the first visit, another scraping should be done after a few days, up to a maximum of four attempts. If none of these attempts are successful, the website is assumed to be inactive. Usually, massive web scraping of thousands of websites may last up to 2 weeks to assure that all scrapable websites responded. Another important issue is how to deal with enterprises that have changed their URL. This can be done by performing an URL search for enterprises that were found to have an inactive URL.

Avoiding out-of-date information on websites (Table 3 issue 3) requires regular visiting and scraping websites. This is the key to obtaining high-quality web data. However, there are several enterprises that provide very limited information on what they actually do. In addition, some websites remain publicly accessible even if the enterprise is no longer operating.

Blocking robots (web scrapers) by the website owner (Table 3 issue 4) is a very important potential obstacle to consider. One solution is informing the website owner about the intention to scrape their website and asking their permission for that, but this is not convenient while scraping thousands of websites. The alternative is to use another supplementary scraping approach, such as a web-browser-based engine, e.g. a headless browser from Chromium or Mozilla Firefox, that might mitigate the criteria used by the website owner and prevent being identified as a robot. The use of these web browser engines should be the same as the typical use of web browsers, i.e. there should be delays between requests and not all the attachment links (e.g. PDFs) should be scraped. Using such an approach will certainly increase the number of the collected websites. In our case study, we observed that adopting this option increased the positive response of about 10% of the websites. It is very important to indicate the robot properly in the user agent variable by including a 'web scraper for official statistics'.

Robots.txt rejection (Table 3 issue 5) is challenging. On the one hand, it is possible to access the data even if the robots.txt denies this type of traffic on the website; we found it was possible in 95% of such cases. On the other hand, we should respect the rules laid down by the website owner in the robot.txt file. However, since the data is used for the production of official statistics, our suggestion is to scrape the data and inform the website owner about this unexpected traffic via the appropriate channels used by the NSI of the country.

Website temporary unavailability (Table 3 issue 6) is also related to the second item on the list of possible reasons for the unavailability of a website, i.e. outdated URLs. However, in this case, we are focusing on the temporary unavailability of a website. This issue can be solved by repeating the requests at another date and time (as mentioned before). If the requests repeatedly fail for a large numbers of websites, it might be helpful to change the IP address, use a VPN or delete cookies.

In many cases, the time stamp is very important when collecting data (Table 3 issue 7). For example, when collecting job advertisements, it is important to have information on the date and time when a specific job advertisement was published. The seventh issue shown in Table 3 illustrates a situation where it is not possible to extract the time stamp from the website. If that is the case, one option is to perform web scraping at a regular basis for a specific period, e.g. daily or weekly, according to the requirements of the research, and to compare the results to see what has been added or removed. However, an easier soultion is to simply use the date of web scraping, withouth worrying when the ad was first published. If the advertisement is on a website, most probably it is valid.

Website duplicates (Table 3 issue 8) come in two different forms. One occurs when a selected number of websites is scraped, in order to, for example, obtain real

estate advertisements, and the same or a very similar advertisement is discovered in the data collected. The second relates to a URL list in which one URL is used by several (different) enterprises or when different URLs redirect users to the same website. In the first of the above-mentioned cases, it is necessary to include the detection of similar items in the data-processing phase. Very similar advertisements should be treated as the same record. In the second case, where a URL is linked to more than one enterprise, it is important to check whether the enterprises with the same link are connected (e.g. branches, etc.). If this is actually so, the results of the website analysis should also be linked to each of these enterprises. When different URLs refer to the same webpage, this usually indicates that an enterprise wants to increase the traffic to its webpages by increasing the chance that the website is found. In this case, it is important to verify if the original URLs are all correctly associated with the enterprise and not with others.

Limited amount of information provided by websites (Table 3 issue 9) can negatively affect the results of web-scraping. Manual check is required to assure that the website is still owned by the company and that the information extracted is all that is available on the webpage. We found that this situation is quite often caused by websites reporting that the domain is either 'for sale' or 'under construction'. If this is the case, these websites are actually inactive and need to be excluded. However, when the website is owned by a company and some (relevant) information is provided, this needs to be included in the subsequent processing steps. One way of dealing with these kinds of websites is to include them in the final estimation process as a separate group (Daas & van der Doef, 2020).

The quality of the relation between an enterprise (sample unit) and its website (Table 3 issue 10) needs consideration when, in some cases, there are errors in the databases resulting in a website not being linked to the appropriate company. It is also possible that the domain has expired due to non-payment and requests are redirected to the website of the service provider. In all these cases, the solution is to check the content of the website and compare it with the data in the business register, i.e. the company's name, address, etc.

Collecting data on enterprises which do not have a website but are included in the sample population (Table 3 issue 11) is only a problem if the data from these enterprises is required. Obtaining all the required information from enterprises without a website may be difficult when a large sample is studied. However, one can attempt to study a smaller population, e.g. a selected type of companies, for which data is available in another source, such as a survey. If this is possible, one could attempt to estimate the total number of companies with no website to get an idea of the size of this part of the population. Such an approach is briefly explained in the study on innovative Dutch companies (Daas & van der Doef, 2020). In this study,

the number of innovative companies without a website was estimated via the units included in the Community Innovation Survey and by contacting a small sample of potential innovative small companies directly. The final estimate of innovative Dutch companies without a website was 0.1%.

4. Discussion and examples

The use of website data in statistical production cannot be overestimated. First of all, all kinds of data on the demand on the labour market, real estate advertisements or price statistics can be downloaded this way at a minimum cost. The cost is predominantly the work of people collecting and processing the data. The essential issue is obtaining a representative (part of a) population to be used in the study. Knowledge of the market and the largest players in the studied field might be helpful. During our extensive work in the area of web scraping, we formulated some helpful recommendations. Firstly, when looking for the most relevant websites to be scraped, do not assume that the biggest are the best. For example, there are numerous websites with job ads, but only a few of them have been stable and reliable over time. This is essential, as time series might be significantly disturbed by including websites with volatile job advertisements in the population. As shown in Table 1, it affects P1 and P2 population sizes.

If voluminous web scraping is necessary, the authors of this study prefer to scrape as many websites as possible. A typical example of massive web scraping is the Online-Based Enterprise Characteristics (OBEC) survey. It has been repeatedly conducted for a selected number of EU countries and the results can be found in the experimental statistical website (European Commission, n.d. c). The difference between the case study described in this paper and the OBEC survey is the population size. In our case study, we used all the websites of Polish enterprises available in the Orbis database. In the OBEC survey, on the other hand, which is conducted for several EU countries, only those websites of the OBEC population were scraped that have been traditionally used in the survey on the ESS enterprises' application of the ICT. The expectation of similar results for both studies independent of the data collection mode is the main motivation for using website data to extract enterprise characteristics.

However, the voluminous web-scraping case study described in this paper demonstrates that when this technique is used, researchers should make allowances for the issues described in Table 3. Here, it actually helps to collect as much data as possible. Additional advantage of massive scraping is that the findings for websites of smaller enterprises (i.e. those with fewer than 10 employees) are included, which is not the case in traditionally-conducted surveys. On the other hand, excluding

enterprises with fewer than 10 employees makes the mitigation of problems much easier. This is why there are some suggestions to limit the population to the most reliable URLs. The same arguments hold for the study of innovative companies (Daas & van der Doef, 2020). In this Dutch study it was shown that i) the traditional survey-based estimate of the number of large innovative companies could be done with web-scraped data only, and that ii) the number of small innovative companies could be determined for the first time.

Slightly less complicated is a study that uses a selected number of websites, like job advertisements. Our experience of working with such data shows that these can be easily included and fulfill the requirements of a traditional survey. Additional advantage of small-scale web scraping, i.e. based on a smaller number of websites, is the fact that there is the possibility to contact all website owners and inform them about scraping their domains. However, the occurrence of duplicates in web-scraped data is a much more complicated problem. For instance, in the case of job advertisements, many providers might add (a set of) the same advertisements, some of which may even be repeated at different locations within the same domain. According to our experience with the OJA data for Poland from the four largest OJA portals, up to 10% of the job advertisements collected were found to be duplicates. Duplicates can be very difficult to detect, because the enterprise to which the job applies sometimes cannot be accurately identified.

5. Conclusions

This paper presents an overview of issues that affect the use of website data for official statistics. Some of these problems are of technical nature, and can be solved relatively easily. However, as our case study, the analysis of the literature and our personal experiences demonstrate, the most challenging problem is related to the selection of a set of websites, as well as the quality of the link between the units and the websites used in the study. The larger the number of websites used, the more serious these issues become. The difficulty is that a certain number of pages and objects needs to be collected to represent the target population. This population may not be known in advance; what is more, it is often determined no sooner than the data has been collected. This particularly concerns enterprises where, based on the Eurostat data, it should be possible to estimate the percentage of firms having a website, but it may not be possible to indicate exactly which of them actually have one. The previously mentioned URL-search approach can be used here to mitigate this problem.

The application of internet robots using search engines provides an opportunity to increase the number of URLs. This increases the sample size, which is very helpful

when conducting research based on websites. External sources provided by third-party companies may also prove helpful for public statistics. The synergistic effect of different URL retrieval methods and additional sources will certainly contribute to creating a list of URLs as complete as possible, and are also likely to enhance the relationship between an enterprise and the accompanying website. However, it should not be forgotten that the Internet is constantly changing and data collected today may differ significantly from the those collected tomorrow. This is relevant for all studies that use web data. Time stamps are essential here.

This enables us to answer the research question posed at the beginning of this paper, namely whether it is possible to collect internet data suitable for the production of official statistics using massive web scraping techniques. The tentative answer is that one source of URLs (a database) may not be enough to conduct reliable massive web scraping surveys. Multiple sources, which need to be maintained, managed and updated, and supplemented by third-party providers (whenever possible), are generally preferable. The case study conducted on one database revealed that nearly 20% of URLs were not accessible, due to non-existing websites or website owners blocking access of the robots.txt file to prevent scraping. Therefore, we suggest using the methods and solutions listed in Table 3 to deal with these issues. When all information is available and the data has been processed, subsequent bias correction methods need to be applied to produce the best possible estimate.

References

Anglin, K. L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. *Journal of Research on Educational Effectiveness*, *12*(4), 685–706. https://doi.org/10.1080/19345747.2019.1654576.

Antonov, O., & Laktionova, O. (2020). Evaluation of Real Estate Market Value in Ukraine Using Web-Scraping. *Galician Economic Journal*, *63*(2), 35–44. https://doi.org/10.33108/galicianvisnyk_tntu2020.02.035.

Ascheri, A., Marconi, G., Meszaros, M., & Reis, F. (2022). Online Job Advertisements for Labour Market Statistics using R. *Romanian Statistical Review*, (1), 3–26. https://www.revistadestatistica.ro/2022/03/online-job-advertisements-for-labour-market-statistics-using-r/.

Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of Gold: Scraping Web Data for Marketing Insights. *Journal of Marketing*, *86*(5), 1–20. https://doi.org/10.1177/00222429221100750.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Inflation Measurement and Research. *Journal of Economic Perspectives*, *30*(2), 151–178. https://doi.org/10.1257/jep.30.2.151.

Daas, P. J. H., & van der Doef, S. (2020). Detecting Innovative Companies via their Website. *Statistical Journal of IAOS*, *36*(4), 1239–1251. https://doi.org/10.3233/SJI-200627.

Daas, P. J. H., Puts, M. J., Buelens, B., & van den Hurk, P. A. M. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, *31*(2), 249–262. https://doi.org/10.1515/jos-2015-0016.

Dogucu, M., & Çetinkaya-Rundel, M. (2020). Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *Journal of Statistics and Data Science Education*, *29*(sup1), 112–122. https://doi.org/10.1080/10691898.2020.1787116.

European Commission. (n.d. a). *ESSNet Big Data I*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/content/essnet-big-data-1_en.

European Commission. (n.d. b). *ESSNet Big Data II*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/essnet-big-data-2_en.

European Commission. (n.d. c). *Experimental big data statistics*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/content/Experimental_big_data_statistics_en.

European Commission (n.d. d). *Web scraping policy*. Retrieved April 21, 2023, from https://cros-legacy.ec.europa.eu/content/item-04-web-scraping-policy_en.

European Commission. (n.d. e). *Trusted Smart Statistics – Web Intelligence Network*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/WIN_en.

European Commission. (2022a). *Deliverable 2.1: WP2 1st Interim Progress Report.* https://cros.ec.europa.eu/system/files/2023-12/wp2_deliverable_2_1_wp2_1st_interim_progress_report_20220331_revision_2.pdf.

European Commission. (2022b). *Report: URL finding methodology*. https://cros-legacy.ec.europa.eu/system/files/20220131_url_finding_methodology.pdf.

Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications*, *13*(3), 144–168. https://doi.org/10.15849/ijasca.211128.11.

Krotov, V., & Tennyson, M. (2018). Research Note: Scraping Financial Data from the Web Using the R Language. *Journal of Emerging Technologies in Accounting*, *15*(1), 169–181. https://doi.org/10.2308/jeta-52063.

Nasiboglu, R., & Akdogan, A. (2020). Estimation of the Second Hand Car Prices from Data Extracted via Web Scraping Techniques. *Journal of Modern Technology & Engineering*, *5*(2), 157–166. http://jomardpublishing.com/UploadFiles/Files/journals/JTME/V5N2/NasibogluR.pdf.

Oancea, B., & Necula, M. (2019). Web scraping techniques for price statistics – the Romanian experience. *Statistical Journal of the IAOS*, *35*(4), 657–667. https://doi.org/10.3233/SJI-190529.

Office for National Statistics. (n.d.). *Web Scraping Policy*. Retrieved August 17, 2022, from https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/webscrapingpolicy.

Orbis. (n.d.). *Overview* [Data set]. Retrieved April 28, 2023, from https://www.bvdinfo.com/en-gb/our-products/data/international/orbis.

Palys, T. (2008). Purposive sampling. In L. M. Given (Ed.), *The Sage Encyclopedia of Qualitative Research Methods, Vol. 2* (pp. 697–698). Sage. https://doi.org/10.4135/9781412963909.

Pegueroles, P., Guerrero, R., Fernández, A., & López, D. (2021). Price's Index through of Web Scraping. *Revista Chilena de Economía y Sociedad*, *15*(1), 32–54. https://rches.utem.cl/wp-content/uploads/sites/8/2022/01/revista-chilena-de-economia-y-sociedad-vol15-n1-2021-Pegueroles-Guerrero-Fernandez-Lopez.pdf.

Polidoro, F., Giannini, R., Lo Conte, R., Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, *31*(2), 165–176. https://doi.org/10.3233/SJI-150901.

Purposive sampling. (n.d.). In *Oxford Dictionary*. Retrieved April 28, 2023, from https://www.oxfordreference.com/display/10.1093/oi/authority.20110810105658510.

Schedlbauer, J., Raptis, G., & Ludwig, B. (2021). Medical informatics labor market analysis using web crawling, web scraping, and text mining. *International Journal of Medical Informatics*, *150*, 1–9. https://doi.org/10.1016/j.ijmedinf.2021.104453.

Wirthmann, A., & Reis, F. (2021). *The Web Intelligence Hub – A tool for integrating web data in Official Statistics*. 63rd ISI World Statistics Congress, Online. https://cros-legacy.ec.europa.eu/sites/default/files/isi_-_web_intelligence_hub_eurostat_paper.pdf.