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.

1 Introduction

These days we use the internet for business, education, shopping, sports, travel, fun and many other tasks. For all these tasks we need data and in many cases we also produce data, some of which is available to others in one way or another. Hence, the web has grown into a huge repository of knowledge and information on economy, wellbeing and life in general. You can hate it or you can love it, such a rich data source cannot be neglected by providers of official statistics. Many statistics offices around the world do indeed show that data retrieved from the web can be used to reduce the response burden, improve or speed up statistical production processes or develop new indicators. This paper builds on these experiences and generalizes the concept of web scraping for official statistics into a more generic framework for use in future projects.

We start with some general definitions:

- A *web source* is any data source that we can observe by executing queries of whatever protocol on the internet. So a web source is not only a website, it also includes ftp sources, the domain name system, mobile versions of websites or streaming media of whatever format. However, although the definition is broad, the majority of this paper will be focused on scraping websites. Within these we focus on the access to *public* web sources.

- *Web scraping* is a term that is used broadly to retrieve data from a web source automatically, that is without human interaction. This is performed by a computer program which is called a

scraper. Similar terms being used for such programs are internet robots, spiders, bots etc. In practice such programs are under control of a statistician and started from a laptop, a program running somewhere in the cloud or on a server. Such a program usually performs some cleaning of the raw internet material and transfers the result in a more or less structured format into the premises of the statistics office for further processing

- We use the tern *web data* throughout this paper to denote the data that has been retrieved by web scraping from one or more web sources. We use the term web data for both raw data as well as data that has been cleaned during web scraping.

This paper is structured from practice to theory. A considerable portion of the paper is reserved to look back on earlier projects to discover successes and pitfalls. We think that only by reflecting on earlier experiences we have enough experience to formulate a more general methodological framework for web scraping in official statistics.

In chapter 2 we summarize and review a number of example web scraping projects around the world. We also briefly highlight the projects executed at statistics Netherlands. For all projects we identify lessons learned. After a brief explanation on technical and legal aspects in chapter 3, the next chapter describes the general phases that we see in almost every web scraping project. In chapter 5 we connect web scraping to survey methodology and in chapter 6 we draw some conclusions.

2 Web scraping experiences

2.1 International experiences

One of the first organisations that proved that data from web sources can be a valuable input for statistics was MIT (Cavallo, 2010). They collected prices from online retailers for a number of years and showed that it was feasible to calculate a price index from this data. At first this was done for a number of South-American countries, which led to some interesting facts on comparing the MIT CPI with the official CPI. For Brazil, Chile and Colombia the scraped indexes captured the main trend in inflation but for Argentina they were not consistent. Later, the process was scaled up for other countries and these days a spin-off company - called PriceStats - collects online prices for many economies worldwide to calculate and sell figures on global inflation trends. These projects were a wakeup call for statistical offices to study the possibilities of online price collection for their own CPI. These days many NSIs, Statistics Netherlands is among them, use web scraping to retrieve online prices in their CPI processes.

Soon after this first project Google started the Google price index, composed of internet data on price offers. It never went into production and there is not much information available these days on the success and challenges of the project. Given the amount of price data Google has access to, we suppose it was not the lack of data that was the problem. More obvious, the construction of a methodologically sound price index from a huge web source of offer prices only could have been challenging. Although we cannot be sure about the exact circumstances in this case, we feel that one lesson to be learned here is that, in order to use web scraped data to produce a statistical indicator, volume is not enough. The methodological design is also important and a challenge.

These early activities using price data from the web inspired the statistical offices from Italy, Germany and The Netherlands to experiment further. ISTAT (Polidoro, 2015) scraped many prices for a well-defined list of consumer electronics products. This approach is different from the big data approaches

from MIT and Google in the aspect that it reuses part of the methodological design, more precise the sample of products chosen, and only replaces the traditional collection process. For this particular set of products the results were promising. More complicated however was the collection of airline fares using the same approach. Apparently successful web scraping in one domain doesn't imply success in another domain.

Destatis (Brunner, 2014) managed to maintain for a longer time a large number of web scrapers for automatic collection of on line prices from web shops and other online retailers in Germany. They gained a lot of experience monitoring price data and concluded that "changes need to be monitored in a targeted manner and a strategy for handling technical losses must be drawn up". At this point it became clear already that there actually is a huge spectrum of different choices that could be made in scraping design and methodology at NSIs, some of which were successful, others needed further research.

An interesting approach in a totally different domain was executed by our colleagues from the Italian institute (ISTAT) (Barcoli, 2016). They used web scraping to retrieve data from the one of the most important data hubs in agritourism. In addition, they scraped many smaller websites from agritourism enterprises. Both sources were used to update and complete the Farm Register. This example shows that data from different web sources can be linked together to improve a statistical register. It is also very interesting in the choice to combine data from both individual websites as well as a so-called aggregator site: a website that organizes content in a certain domain and sometimes also adds structure in categorizing the content (for example in geographical regions, type of activity or other metadata). Generally speaking this poses the question which of the approaches should be preferred. To rely on the original source or to use a maybe better organized data hub with possibly extra metadata. We will come back to this later.

The Office of National Statistics (ONS) and other countries looked into using web data on Job vacancies for statistics (Swier, 2016). This appears to be a tough area for producing official statistics based on web data. Generally speaking, data can be retrieved from job portals or enterprises' sites. One of the tricky aspects is the methodological design and the coverage of the vacancies advertised on line in particular. To make it even more complex the coverage varies heavily depending on the type of business activity and maybe also other variables such as the number of employees. We conclude that these issues are important to pay attention to.

Another subject that had the attention of the ONS recently was an attempt to measure sustainability reporting via web sources (Sozzi, 2017). ONS experimented with the use of scraping and text mining on web sites of UK companies to detect sustainability information automatically. The conclusions state that "it is possible to discern the number of companies publishing sustainability information via scraping of their websites". Also, they claim that the approach could be repeated in other countries. This is a clear example of two key principles: 1) the use of a focused scraper to navigate to a specific part of a web source (here a page containing sustainability information) and 2) the power of processing unstructured information from the web by text mining and natural language processing techniques for use in official statistics. We think that both concepts are generally applicable in web scraping.

2.2 Experiences at Statistics Netherlands

The early activities of Statistics Netherlands in web scraping for statistics started in 2009 (Hoekstra, 2012). The first projects focused on the technical feasibility of scraping data from the web. One of the first cases was the daily collection of online prices of unmanned petrol stations. The conclusion was that this was feasible and we started collecting data on fuel prices regularly. However soon it appeared

that similar data, but then for *all* fuel station brands, could be retrieved directly, i.e. not by scraping, from a lease company tracking fuel prices for their own purpose. It was decided to use that source and this is still the case up to now. An early conclusion that could be drawn from this case was that a web source, although easy accessible and very up to date, sometimes loses competition from direct access to similar data from a data owner[1].

Inspired by this early work, Statistics Netherlands launched a research project to imitate the existing manual collection of airline ticket prices from airline companies' web sites. Six airline companies were scraped daily by a scraper from Statistics Netherlands and two additional ones developed by external companies. They all used different scraping technologies. Although these scrapers did the job well and their data was comparable, it was also concluded that imitating the manual collection process, designed mainly for human collection, was in this case not the most efficient way to proceed. The lesson we learned was that we should not just imitate existing collection processes; we should redesign the collection process based on the features of the medium at hand.

Looking back, it is interesting to see that in our early research on web scraping (Hoekstra, 2012) we concluded that "methodological applicability will have to be assessed on a case by case basis". This has been the case for the past years. It is only in this paper that we feel we have enough experience to try to formulate a more general methodological framework for web scraping in official statistics.

From 2012 onwards we have been increasing the amount of prices collected from web shops mainly for clothing. These days (2018) about 20 scrapers are active to collect about 0.5 Million prices per day which are classified into an elementary aggregate level classification and merged into the CPI using a methodology specifically designed for this (Griffioen and ten Bosch, 2016). Because of the volume of the data and the scraping method applied, we call this type of scraping *bulk scraping* (ten Bosch and Windmeijer, 2014).

In addition to bulk scraping at Statistics Netherlands we explored a totally different approach, which is known as *computer assisted price collection* (ten Bosch and Windmeijer, 2014). Instead of retrieving all offer prices from web sources in this approach we check for *price changes* of well-defined products. For this a dedicated tool has been developed for price analysts: the robot tool[2]. This approach appeared to be particularly useful for web sources that contain few prices that do not change too often such as cinema prices, driving lessons and restaurant menus. It would be too costly to develop a scraper for each of the web sources and the use of the generic robot tool speeds up collection in a flexible way. Comparing this approach with bulk scraping we learned that both have their pros and cons and should be applied in the right context. In the next chapters we will try to give guidelines what to apply when.

Another domain where we explored web sources for statistics is the Dutch real estate market (ten Bosch and Windmeijer, 2014). We analysed about 30 Dutch property sites to see which content was available, what the volume was and what the variables available were. Based on this analysis we started scraping six sites daily. The analysis of the data showed that there was considerable overlap in content and also that one of the sites was usually (but not always) one day ahead reflecting changes. We concluded that this one was leading in content but also that at least one of the others was also statistically relevant. The scrapers were maintained for two years and then replaced with a direct feed from one of the (non-leading but very complete) data providers. We concluded that the scraping

[1] A similar example can be found in the consideration to use scanner or web data for price statistics. If possible, scanner data is preferred. One additional reason is that scanner data provides quantities, expenditures and transaction prices, where web data has offer prices only. However, web sources often contain much more detailed information about the items on sale.

[2] https://github.com/SNStatComp/RobotTool

exercise was useful to explore the real estate market initially and that the experimental scrapers could be taken into production any time in case we would like to replace the direct feed by web data..

A totally different scraping project at Statistics Netherlands was performed for the international EuroGroups Register (EGR) project (van den Heuvel, 2013). Wikipedia was explored as a source for collecting relevant information for the maintenance of the register of internationally operating enterprises contained in the register. Starting from a list of enterprises, in total more than 41 000 articles from Wikipedia (English, French and German version) with basic information was retrieved, cleaned / normalised and delivered to the project. One interesting observation in this project was that the Google search engine performed better finding the right article on Wikipedia than the Wikipedia search engine itself. This we have seen in more cases, which led to the observation that a general search engine can be used to discover, and also partly retrieve[3], data from another web source. We used this "*indirect query method*" in other projects[4] as well.

It is sometimes difficult to classify enterprise activity. In some cases statisticians derive the activity type from the products being produced and they enter product names and serial numbers in a general search engine such as Google. We facilitated this using an automated access to a search engine, but that is not so special, the interesting observation here is that the internet apparently contains detailed information to be used for mapping[5] individual statistical units into statistical classifications. What we learned from this is that we should not only think of the internet as a way to provide us primary data, web sources can provide us useful *metadata* for statistics as well.

Enterprises, both commercial and governmental organisations, produce annual reports containing economic figures. These figures are of particular interest for accounting statistics. The process of retrieving, maintaining and parsing these annual reports is clearly an area where web scraping and text mining could help. The scraping design we developed includes an iterative approach where information on the location of the annual report is maintained over years so that we would build up an increasing knowledge base of enterprises and annual report locations and contents. However, this approach was not successful. One of the reasons could be that it takes time. We learned that such an approach needs to be simple, explainable and deliver within a reasonable time to be accepted.

Social media are a useful source for studies on human or economic activities. However there is usually a strong bias as the social media population differs in many aspects from the normal population. In 2016 Statistics Netherlands performed a selectivity study (Daas, 2016) using both twitter as well as LinkedIn as a source of data to estimate this bias. This web scraping exercise showed the power of combining different web sources, which both have their strengths and weaknesses, to derive auxiliary information to be used in official statistics.

More recently Statistics Netherlands worked together with the NSIs from Italy, Poland, Sweden, UK and Bulgaria on different use cases for retrieving data on enterprises via web scraping (ESSnet WP2 deliverable 2, 2017). Since the business register does usually not contain the URL of an enterprise the first step in this process is to find it. Search technology in combination with focused[6] crawling is used to find candidate URLs for these enterprises. Machine learning is applied to predict whether a URL

[3] A search engine usually produces a small text from the site, called a *snippet*, which we can use for this.
[4] There might be other reasons to turn to an indirect query / scraping approach such as practical reasons (ease, speed) or legal reasons (a scraper does not visit the site itself only the search engine).
[5] Knowing that web sources and search engine technologies continuously change, one could have worries about the stability of this mapping used for statistics. The opposite is also true. Official statistics not following the ever changing digital world in their (static) classifications could be worrying as well. We leave this for further discussion.
[6] A focused crawler visits only parts of a web source, for example the contact page of an enterprise.

does indeed belong to the enterprise of interest. Other use cases are the detection of Ecommerce activities, Social media presence and use. What we learned from these projects is that it does make sense to start web scraping from the – possibly incomplete - data that we already have in our statistical databases, in this case the business register.

3 Technical and legal aspects of web scraping

A necessary precondition to the use of web scraping in official statistics is mastering the technical and legal aspects.

3.1 Technical aspects

Although interesting and crucial for any successful web scraping project we choose not to describe the technical aspects of web scraping in this paper. It might fill up a paper itself and to make it even worse fast changing technologies make such a detailed description out of date soon.

We have seen the internet changing from a mainly static medium in the nineties to a more interactive communication mechanism in the first decade of this century to the dynamic interconnected network of web sources that it is now and it will keep changing. In 10 years of scraping we changed from Perl to a tool named Djuggler to the programming language R to a combination of Python and JavaScript / Node.js that we use now. Scraping concepts differ from reading HTML to parsing XML or JSON hidden in web pages to querying REST interfaces that connect to backend directly. This all depends on the web source of interest. One has to continuously monitor the validity and usefulness of the tools chosen and the scraping mechanism applied. However the main concept of automatically reading data from web sources has been stable. For a more detailed description of this concept we refer to our earlier work (ten Bosch and Windmeijer, 2014).

We conclude here that although web scraping requires a vast amount of technical knowledge and experience it can be managed successfully by people with the right knowledge and experience choosing the right tools based on the latest technologies.

3.2 Legal aspects

Legal aspects of web scraping might differ slightly per country. Generally speaking one has to meet the national statistical law, legislation on database protection, intellectual property rights, privacy protection and netiquette (etiquette on internet). It would not be feasible to discuss them here all. A thorough overview of the aspects involved in different countries can be found in (Stateva ea. 2017).

With respect to netiquette we briefly mention the most important aspects, being to identify your scraper using the user-agent string, to not overload servers using some idle time between requests and run scrapers at night, to respect the robots exclusion protocol and to inform website owners of using their data in production if feasible. We do not see any problems in web scraping of public web sources as long as these principles, collectively called the principles of transparency and unobtrusiveness, are taken into account.

4 Phases of setting up web scraping for official statistics

In this chapter we build on the experiences achieved in the example projects presented in the previous chapter to identify the general aspects important for setting up a successful web scraping project. The

goal of this exercise is to create a more generic reference framework for web scraping for official statistics. We do this by making a distinction between three phases[7]:

- the *site analysis* phase
- the *data analysis* and *design* phase
- the *production* phase

In the site-analysis phase a web source or multiple web sources is / are examined for its technical features from the viewpoint of the web scraping and data availability. In the analysis phase a test data stream is set up which is analysed. Depending on the data source this phase could take up a few months to a few years. The production phase is when the data source is used in the production of official statistics, whatever way. Obviously, if the analysis phase does not have a sufficiently positive outcome, a data source will not reach production.

Note that in some cases, for example for a research project where data is collected only once to study a specific phenomenon, there is no need to enter the production phase. However, the concepts mentioned in the two other phases still apply.

For each of the phases we list some of the key characteristics being examined:

4.1 The site analysis phase

In the site analysis phase a scraping expert examines the navigation and data characteristics of a web source. Among the questions to be answered are:

- Programmability: is there an application programming interface (API) available. If so, what are its opportunities and limitations? If not, is there another way to query the site, for example using a REST query being observed, or indirectly via a general search engine?
- What is the estimated amount of items on the site?
- What is the estimated volatility of the data on the web source?
- Legal: Is there a site exclusion protocol or meta tags to be respected. What is the legal status of the content on the site, are there any specific requirements?
- Is this content original or is it copied from another source (or vice versa). If so, what web source(s) would be the preferred supplier?
- Is there another web source that offers access to the targeted web source(s) in a more coordinated way? Could it be useful to use that one instead?
- Level of detail: should the source be scraped in detail or only the front page and/or only to a certain depth. Only the product list or also every product item page? This depends on the availability of variables on each level.
- Menu structure: do the menu items contain valuable metadata? If so, could be use it to traverse the site? Do certain categories cover the whole site? Which categories do we want to observe? Is the menu structure strictly hierarchical?
- Business model: Is the web source controlled by a well-known business firm, a group of enthusiasts collectively working on data, or the result of a one-time research exercise? That influences our decision to use it in the long run.

[7] The aim of this paper is to sketch a generic framework for future web scraping projects. More phases could be identified, but to keep things simple and recognizable we restrict ourselves here to these three main phases.

The results of the site analysis phase are described in a site analysis report, which is used to decide upon setting up a data stream for the candidate web source.

4.2 The data analysis and design phase

In the data analysis and design phase a data scientist analyses the data being retrieved in context to the target variables of the statistics, or possibilities for new indicators. Questions to be answered are:

- What is the stability of the scraping mechanism chosen?
- What is the observed amount of items on the site?
- What is the observed volatility of the data over time?
- How plausible are the variables retrieved and how do they relate to statistical units?
- How stable is the metadata retrieved. Do the categories expressed by the web source change often? Is there a pattern?
- Identifiability: can we link the data retrieved to one of our statistical object-types (an enterprise, a product, a building, a region, a farm, a household etc.) in our statistical databases?
- Combinability: can we link multiple observations from the same or different web sources at different times together?
- What is the statistical design to be applied if this web source is used in production? What is the role of the web data in the statistical process? Replacement of survey data, quality improvements, new indicators? See also next chapter.
- If this web source is to be used in production, what would be the ideal scraping frequency? Note that to answer this question one has to analyse the volatility first and this can only be done using a high frequency data stream. In many cases the frequency of a production stream can be lower than the frequency of the analysis stream.

The distinction between the site and data analysis phase is not always completely strict. Some of the questions in the site analysis phase might better be answered using the data retrieved and vice versa.

Before going into production it is advisable to, whenever possible, inform site owners. Only if a web source consists of many different domains, such as a crawler the hops from one site to the other, it is not feasible to do so and the scrapers must rely on their transparent and non-obtrusive way of working, see earlier.

4.3 The production phase

One of the key characteristics of web scraping is that even in de production phase things can change. One has to organize[8] for that, for example:

- Monitoring the availability of the data stream
- Monitoring the data for implausible changes in variables
- Organising a team of specialists that can act quickly upon incidents in the above two points
- Once in a while, for instance every year, re-examining the web source(s) and data stream for unusual data or metadata changes that were not noticed during normal monitoring
- Maintaining the relationship with site owners

[8] Although not mentioned explicitly, the monitoring of data streams is also important in the data analysis phase. Often we need longitudinal data for analysis that is reasonably stable for a reasonable time.

5 Web scraping and survey methodology

In this chapter we take a more formal look at web scraping for official statistics from the viewpoint of survey methodology. For this, we imagine that we have a phenomenon that we could observe by a classical survey approach but also by a web-based scraping approach. For simplicity we imagine that we want to predict one target variable for a population of statistical units[9] using either approach. This population could be for instance a population of persons, enterprises, houses, products etc.

In a classical survey approach (Groves, 2009) we would have a population register, or at least something we could use for that, from which we would draw a sample of statistical units that minimizes the sampling error. We would then design an approach to collect the data for these units, e.g. a questionnaire in whatever mode. Based on this microdata we would then infer a prediction about the target variable for the whole population. This is of course somewhat oversimplified, but we keep it simple to be able to position the use of new web sources in this framework.

We do not have such a methodological framework for using web sources in official statistics yet. Although the number of different use cases for web data in official statistics is sheer endless, we try to find some common patterns from the examples presented earlier. One of these common patterns is that the answers to the questions in the site analysis and data and design phases can make a big difference in the way the web data can be used. For example the possibility to query a web source *for a specific statistical unit*, in other words to query for a specific unit that we have in our statistical databases, makes a difference for our statistical design. Also the *coverage* of the web sources with the population of interest is very important to consider. Further, the *selectivity* of the web sources relative to that of the target population plays an important role. Roughly speaking, the coverage and selectivity of the web sources contribute to the (in)accuracy of final estimates of output based on the web sources.

Based on these observations we formulate a general workflow that could in principle be applied to any web source:

1. The first step is to answer the question whether the web source supports queries for statistical units from our population register. More precisely, does the web source[10] allow us to issue a web query with an identifier from our administrative source? If so, we can use a probability sample to query for data on this specific population element and predict the target variable for the whole population in the classical way, taking coverage and selectivity in consideration. It is important to realise that the web source may not be available for all units in the population. The total sample may therefore be split into two parts: a part for which the web source is available and a part for which this is not possible. For the latter part we could either directly collect the target variable for instance by a new questionnaire or maybe there is an already existing questionnaire that contains information on the target variable. The web source is then used to obtain estimates for small areas in the population, but these estimates may be biased. The survey part could be used to correct or to mitigate the bias, for instance by using a combined estimator. Note that using a probability sample on a web source supporting identifier queries is a *possibility*, not a *necessity*. One could decide not to do that, for many reasons, and proceed as if the answer on the first question was a no.

[9] We use the word statistical unit to denote one population element, in line with the Eurostat glossary: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Statistical_unit. More precise would be the use of the word observation unit, but we feel that this is less often used in this context.

[10] Note that one web source may contain many statistical units. The question is whether the data for one statistical unit can be retrieved by a web query using a known identifier for this unit.

2. If querying for individual statistical units is not possible, or we choose not to do that, we can try to identify such units from the set of observation values in the observations contained in the web scraped data. This is a step that is executed very often in web scraping. This is obvious, it is always necessary to link the statistical world to the virtual (web) world. If linking these worlds together on the level of statistical units is possible, then we can calculate the coverage. If this is large enough we could decide for a probability sample with a bias correction. A bias correction requires the availability of auxiliary variables for all units in the population. These auxiliary variables can only be used as a means for bias correction when their values are related to the probability that the (scraped) data are available and when the values of the auxiliary variables are correlated with the values of the target variable(s). If so, one could correct this bias by means of a pseudo-design based estimator (Buelens et al, 2018), for instance by using a post-stratified estimator (Särndal et al. 1992). If coverage is too small or we decide not to do a sample, we could proceed to the next step.

3. If we cannot link the scraped data to our statistical units on the microdata level, we can try to link the data on a higher aggregated level into official statistics. A typical example is the clustering of products and prices in web scraped data into an elementary aggregate (EA) classification in price statistics retrieved by bulk scraping for calculating the CPI.
   Another possibility might be to derive auxiliary variables from the scraped data that are also available for all units in the population. For instance, for social media data one could try to derive characteristics like whether it concerns 'a person or a business' and in case of a person, its gender, age-class and position in the household . This derivation of characteristics is also referred to as profiling (Daas and Burger, 2015). Subsequently one could apply a post-stratified estimator.
   Further, one could use sample matching (Rivers, 2007; Rivers and Bailey, 2009). In sample matching one draws a random sample from the population, for which a number of auxiliary variables are available (but those characteristics are not necessarily used in drawing the sample). Next, one synthetically links units in the scraped data to units in the sample based on a set of common auxiliary variables. When the auxiliary variables that are used in sampling matching are available for all units of the population, sample matching will give nearly the same result as post-stratification (See also Bethlehem, 2014), whereas the latter is simpler. Sample matching might be useful when the sample and the scraped data (also) contain common variables that are only available for units in the sample and in the scraped data but not for the whole population. For instance one has derived the variables gender, age class and position in household from the scraped data, but one does not have these variables for all units in the population, but only a sample of units.

4. If the aggregation / clustering of web data into an existing aggregate level is not possible, and also reweighting based on auxiliary variables (through post-stratification or sample matching) is not possible then no unbiased inference can be made on the target variable of the population of interest. However, in some cases, if a strong correlation is identified between the scraped variables and a target parameter published in official statistics, one could argue that the web data could be used as a fast beta indicator for that target parameter. There are successful examples of such approach (Daas et al, 2014). Of course one has to be careful here, a correlation detected on data on the past doesn't prove that it holds in future, as the Google Flu project has proven (Lazer et al. 2014). In cases where a beta indicator is always followed by a slower but traditionally sound statistical indicator, this is probably not a problem. Of course,

in this situation it is important that users are informed about the methodology used for the beta indicator.

The workflow described in the previous steps is shown graphically in figure 1.

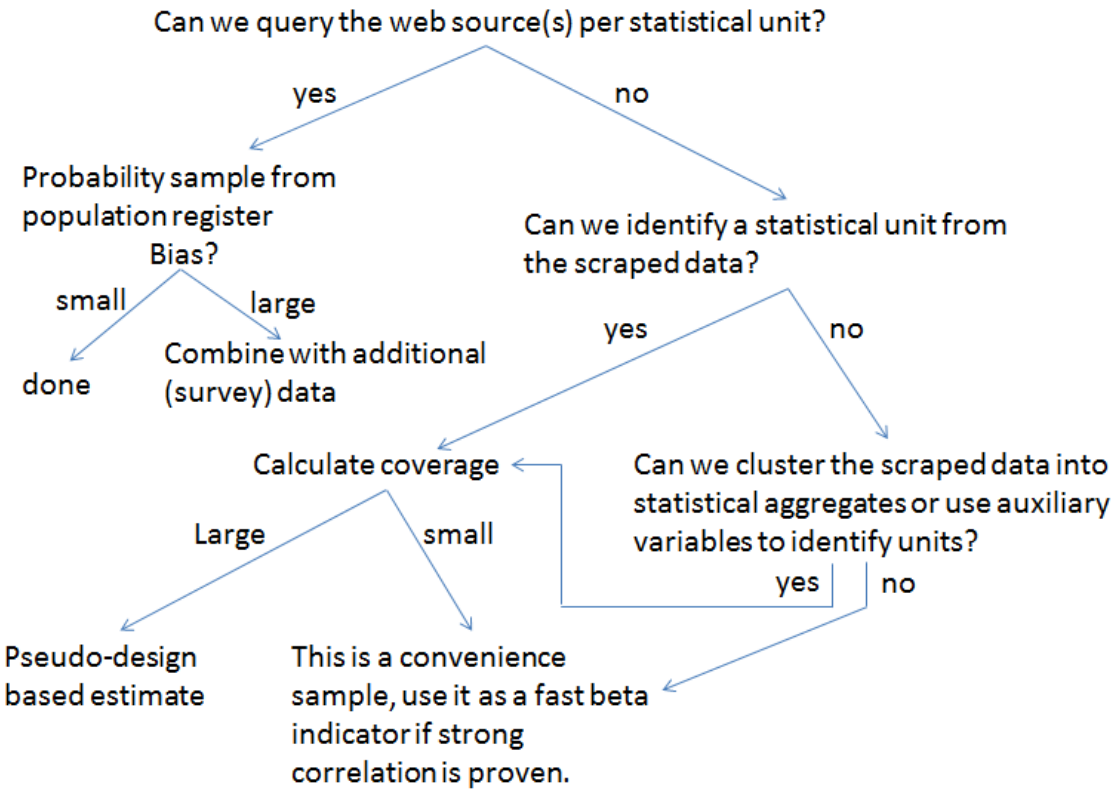

Figure 1: web scraping in the context of survey design.

6 Conclusions

In this paper we looked back on a considerable amount of cases where web scraping was applied in the area of official statistics. We concluded that this field is of growing importance for national statistics institutes. Some of the more prominent use cases are to reduce the response burden, improve or speed up statistical production processes or develop new indicators. But we also signalled that web scraping is used in other circumstances such as to explore background variables, to retrieve metadata or to characterise (sub) populations.

We quickly mentioned the technical and legal aspects of web scraping, which have been described earlier more extensively and concluded that these are, in our view, manageable and solvable in practice.

We identified 3 phases in web scraping which have their own challenges to be attacked: the *site analysis phase*, the *data analysis and design phase* and the *production phase*. In each of these phases we mentioned a number of activities to be carried out and questions to be answered before going to the next phase (if at all). Although probably not complete, we hope that these phase-specific list of questions can help others to structure any new web scraping project.

Based on the reflections on earlier web scraping projects we formulated a first general methodological framework for web scraping in official statistics. One key element in this framework is the ability to *query* a web source per statistical unit or to *link* the web data to statistical units already available in population registers of any kind. This makes a difference for the potential use of the web source in official statistics. If this is the case, some traditional methodological methods can be applied to make sound predictions on the target population. If this is not the case other methods should be considered which were briefly explored and described. All in all we hope this is a first step to the definition of a more general framework applying web sources in official statistics.

7 Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors are grateful to the colleagues from Statistics Netherlands and fellow National Statistical Offices for their work in this area which we used in this paper.

References

Barcaroli G, et al. ISTAT Farm Register: *Data Collection by Using Web Scraping for Agritourism Farms*, ISTAT, ICASVII, Rome 2016

Barcaroli G., Scannapieco M., Scarnò M., Summa D. (2015a). *Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies*. In NTTS 2015

Betlehem, J.G. (2014). *Solving the non-response problem with sample matching?* CBS discussion paper 2014-04.

ten Bosch, O., Windmeijer, D., *On the Use of Internet Robots for official Statistics*, UNECE MSIS conference, Dublin, 2014

Cavallo, A., 2010 *Scraped Data and Sticky Prices*, MIT Sloan, December 28, 2010

Buelens, B., J. Burger & J. van den Brakel (2018). *Comparing inference methods for non-probability samples*. International Statistical Review doi:10.1111/insr.12253.

Brunner, K. , *Automated price collection via the internet*, DESTATIS, 2014

Daas P., Burger J., Le Q. ten Bosch O., Puts M., *Profiling of Twitter users: a big data selectivity study*, Discussion paper Statistics Netherlands, https://www.cbs.nl/en-gb/background/2016/21/profiling-of-twitter-users-a-big-data-selectivity-study

Daas, P. and J. Burger (2015). *Profiling big data to assess their selectivity*. Paper presented at the conference New Techniques and Technologies for Statistics (NTTS), Brussels 2015.

Daas, P. e.a., *Social media sentiment and consumer confidence*, ECB Statistics paper series, No 5 / sep. 2014

Griffioen, A.R., ten Bosch, O. (2016). *On the use of internet data for the Dutch CPI*. Paper presented at the UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices, 2-4 May 2016, Geneva, Switzerland.

Groves R., e.a., *Survey Methodology*, 2nd edition, July 2009

Van den Heuvel, G. *Wikipedia as a source of Business Data*, Internal report Statistics Netherlands, delevered to the EGR project: https://ec.europa.eu/eurostat/cros/content/egr-0_en, 2013

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: Journal of the International Association for Official Statistics 28 (3-4). pp. 99-111.

Lazer, D, e.a., *The Parable of Google Flu: Traps in Big Data Analysis*, Science 2014, Vol. 343, Issue 6176, pp. 1203-1205, DOI: 10.1126/science.1248506

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 (2015) 165–176

Rivers, D. (2007), Sampling for Web Surveys. Paper presented at the Joint Statistical Meetings, Section on Survey Research Methods, Salt Lake City, Utah.Rivers, D & Bailey, D. (2009), *Inference from Matched Samples in the 2008 U.S. National Elections*. Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, Florida.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling* New York: Springer-Verlag

Swier, N., *Webscraping for Job Vacancy Statistics*, Eurostat Conference on Social Statistics, 2016

Sozzi, A. (ONS), *Measuring Sustainability Reporting using Web Scraping and Natural Language Processing*, NTTS, 2017

Stateva G., ten Bosch O., Maslankowski J., Righi A., Sannapieco M., Greenaway M., Swier N., Jansson I., *Legas aspects related to Web scraping of Enterprise Web Sites*, ESSnet Big Data Work Package 2, 2017.