Conference on New Techniques and Technologies for Statistics 2023, 6 - 10 March 2023, Brussels, Belgium

Business register improvements: a balance between search, scrape and 3rd party web data

Olav ten Bosch[1], Arnout van Delden, Nick de Wolf
Statistics Netherlands

Keywords: Business registers, web scraping, search engines, machine learning, NACE

1. INTRODUCTION

These days the web is an integral part of life. We use it in our work, for communication and interaction, to compare, buy and sell products, to plan holidays, to order food and many more things. For enterprises, the web is often much more than a communication channel. Depending on the activities they perform the web may play a role in advertising, in managing their logistics, in customer care, or in recruiting new employees. Hence the web is full of digital traces from economy that may help National Statistical Institutes (NSIs) in making official statistics.

NSIs typically maintain a statistical business register (SBR) of enterprises. It contains the statistical units comprising the enterprise population and variables such as number of employees (size class), location and type of economic activity (NACE code). It also contains the relation between enterprises, other statistical units and administrative units such as legal units. It is used as a frame for drawing samples for surveys and as the backbone for statistics on economy. Using the web as a data source the SBR can be improved with better, more detailed or new information that is difficult to grasp in a more traditional way.

Figure 1 contains a high-level view of this concept. Depending on country characteristics the proportion of legal units for which the website(s) are known may vary heavily. In most cases the missing URLs have to be found in the so-called URL finding phase (A). In the next phase (B) statistical variables can be derived or improved from web data. In both phases websites, 3rd party scraped data or other web data sources can be used. In this paper we list various ways of using web data as examined in the international Web Intelligence Network (WIN) project, WP 3 (new use cases), UC5 (business register quality enhancements). We give some examples, and list the advantages and challenges. The main focus is on phase A and we will dive a bit deeper into a practical case of using 3rd party scraped data.

[1] Corresponding author: o.tenbosch@cbs.nl

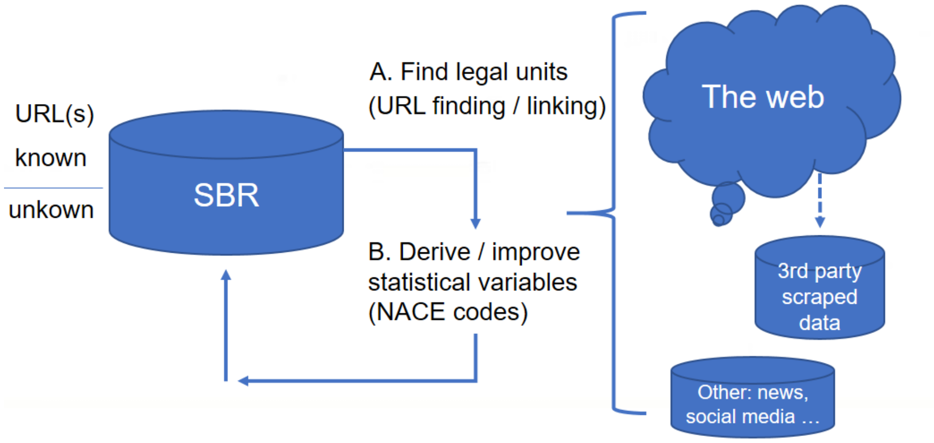

Figure 1: General view on business register improvements from web data sources

2. Search

Search engines are used to find URLs for legal units or to verify known URLs. A search query is executed containing the name of the unit, possibly augmented with identifying information such as the municipality or a chamber of commerce (COC) number. This needs to be automated, so a (paid) search application programming interface (API) might be a better mechanism than interpreting the human readable search result pages, however both approaches are possible. Search results have to be interpreted to select the best match. The snippet – a short textual extract of the results page – can be used to do this, or alternatively, URLs returned by the search engine can be scraped to select the most relevant hits. This however involves an extra scraping step.

In countries where it is obligatory to mention identification numbers on enterprise websites these can be used for direct exact linkage. When such legislation is absent, machine learning techniques have proven to be useful for selecting relevant search results. Based on a labelled training set of valid and invalid search hits a model can be trained that catches the search engine behaviour for that particular search engine. The set of legal units in the SBR with known URL can serve as a training set. Since search engines evolve over time the model used has to be retrained periodically.

Search engines should be used with care as identifying information contained in the query could be identified in web server logs. This phenomenon is known as *search engine leakage*. One can reduce the risk by carefully designing the queries, spreading them across different search engines or by making a non-disclosure agreement with a search provider. Search engine leakage is to be taken seriously but manageable. A more in-depth discussion of using search engines for URL discovery can be found in [1] and [2].

3. Scrape

Scraping websites that belong to legal units in the SBR can only be done if the URL(s) for that unit is known or discovered by search or linking to other sources. If that is the case scraping is typically done via the *generic scraping* concept, for which unlike *specific scraping*, no prior knowledge on the structure of that site is available. Generic scraping typically starts at the home page and recursively visits deeper pages up to a certain maximum depth. Decisions have to be taken whether to store the complete website, only the texts, or the derived data (or all).

A *focused scraper* does not follow all links but gives priority to those that are expected to contain the most valuable information for the task at hand. For example to detect economic activity, a focussed scraper might give priority to the 'about us' page.

It is important to check that the website visited does match the legal unit at hand. The existence of identifying information on the site is crucial for this task. National legislation might force enterprises to put such information on their website, such as Chamber of Commerce or tax id, but this is not always the case. Special care should be given to many-to-many relations between legal units and websites. Depending on the business activity an enterprise might run many different websites. Contrary, small businesses might not have their own website and use a 'business collection' page to advertise their services.

An example of a more extensive discussion of scraping for statistics can be found in [3].

4. 3RD PARTY WEB DATA

Web data collected by 3rd parties can be a useful additional input. Re-using such data saves resources. On the other hand, a (paid) agreement has to be made and the dependence on the 3rd party has to be managed. This is feasible only if the added value of the data is considerable. An example of a 3rd party collecting web data is the company DataProvider (DP). For over 2 years they provide Statistics Netherlands with a monthly DP dataset with URLs of Dutch businesses and additional variables. This data has been linked to the SBR using contact variables such as COC number, domain, email, zip code and phone numbers. Contact information might be missing on websites or in the SBR. Figure 1 shows the gaps in the DP data (left panel), and the corresponding SBR linkage variables (right panel).

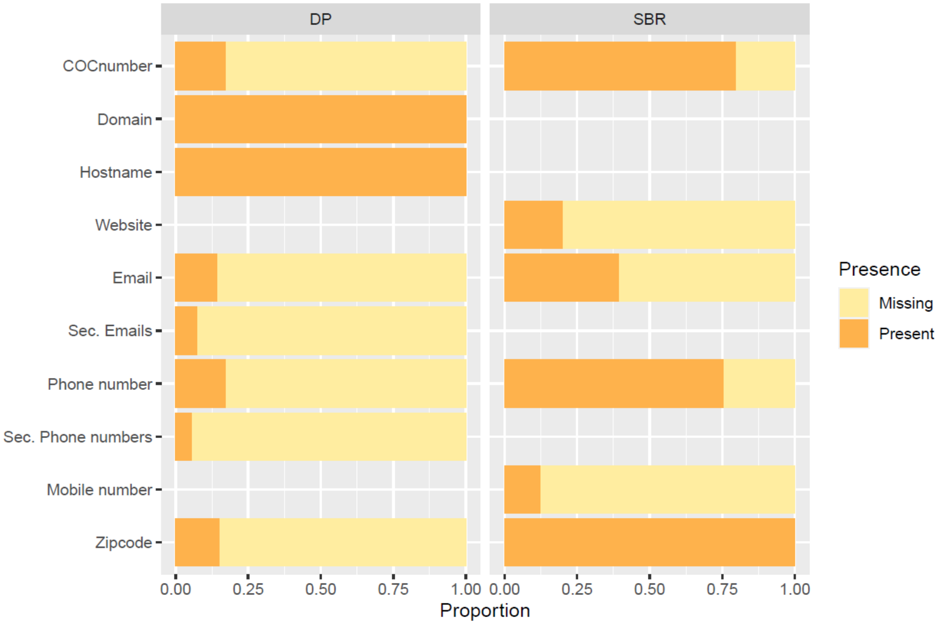

Figure 2: Completeness of linkage variables in DP data and SBR.

Note that the link between DP data and legal units can be one-to-one (1:1), one-to-many (1:n), many-to-one (n:1) or many-to-many (m:n). This makes it a complex task. Table 1 summarizes the results. At 75% linkage probability we have for 528 780 of the 4 630 836 legal units in the SBR (11%) a one-to-one match with the DP data. Moreover for 111 904 + 4 863 legal units (2.5%) there are multiple DP hits that may help improve SBR data. These results can probably be improved if the linking strategy is further refined. The value of the additional variables is still to be examined.

Table 1. Nr. of legal units (LUs) by linkage cardinality at 75% linkage probability

	#URLs in DP			
# LUs	2+ (n)	1	0	Total
2+ (m)	4863	27935	3957354	4630836
1	111904	528780		
0	5057922		X	X

The analysis shows that in this case for about 14.5% of the legal units a URL could be deduced from 3rd party data, which indicates that this approach is valuable. More information on linking all kinds of data (incl. web data) to a SBR can be found in [4].

5. Other web data sources

One of the other data sources that have been used in this context is the internet *Domain Name System (DNS).* This register of domain names and IP addresses is present in all countries and could be useful to deduce domain ownership. However, the degree of openness of this data varies per country and domain. Another data source of interest are *news and social media* messages. There might be some descriptive value in a press release or a social media message about the real activities performed by a legal unit. Yet another data source could be *job ads*. Enterprises typically put short descriptions of their main tasks and activities in it and since these texts are created explicitly for this goal by the company itself, they might be a valuable extra information, if the job ad can be linked to a legal unit.

6. Deriving statistical variables

The scraped data from websites, 3rd party data or additional sources can be used to supplement the SBR or to derive or improve statistical variables. An example of the first is to add email addresses or phone numbers found on websites to the SBR. An example of the second is detecting economic activity (NACE codes) from website texts. The latter requires interpretation of raw texts and usually involves natural language processing (NLP) and machine learning techniques. Other statistical variables that have been derived from web data are, degree of innovativeness, degree of sustainability, operating a web shop or not, or belonging to the platform economy. For a more detailed example on deriving statistical variables from website texts we refer to [5].

7. Wrap-up

In this paper we presented a high-level view on business register improvements using web data.

URL finding concerns the discovery of URLs of legal units that do not have their website registered in the SBR. Search engines can very well be used for this. The textual paragraphs of the search results (snippets) can be used in combination with machine learning techniques to select valid hits. Alternatively a scrape on the found URLs can be performed. Search engines must be used with care, but the risk of search engine leakage is manageable. Once URLs for legal units are known in the SBR or found via URL finding, statistical variables, such as NACE code, can be derived. This involves scraping and interpreting the results using NLP and machine learning techniques. Third party data can be used as an alternative to scraping by the NSI. Other web data sources such as DNS, news or job ads can also be used. The best approach to improve a business register from web data is probably a mix of web data sources and techniques that best meets the country specific situation in the SBR and in the national web.

REFERENCES

[1] A. van Delden, D. Windmeijer, O. ten Bosch, *Searching for business websites*, CBS Discussion paper, Dec. 2019, Searching for business websites (cbs.nl)

[2] H. Kühnemann, et. al. *Report: URL finding methodology*, WIN project, 2022-01-31 Report: URL finding methodology (europa.eu)

[3] G. Barcaroli, et. al *Use of web scraping and text mining techniques in the Istat survey on "Information and Communication Technology in enterprises*, Qconf, Wien 2014

[4] L. Ryan et. al., *An SBR spine as a new approach to support data integration and firm-level data linking*, IAOS 36 (2020) pp. 767–774, doi/10.3233/SJI-200640

[5] Daas, P.J.H., van der Doef, S. (2020) *Detecting Innovative Companies via their Website*. IAOS 36(4), pp. 1239-1251, doi/10.3233/SJI-200627