Distr.
GENERAL

Working Paper
13 February 2014

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2014)
(Dublin, Ireland and Manila, Philippines 14-16 April 2014)

Topic (iii): Innovation

ON THE USE OF INTERNET ROBOTS FOR OFFICIAL STATISTICS

Prepared by Olav ten Bosch and Dick Windmeijer, Statistics Netherlands, the Netherlands

I. Introduction

1. This paper reports on activities relating to the use of so called *internet robots* for official statistics at Statistics Netherlands in recent years. In short, internet robots are programs that imitate operations performed by a human visiting websites via a browser to store interesting data from these sites. Starting with some early prototypes back in 2009, Statistics Netherlands has experimented with this new method of data collection in various projects. Now, in 2014, we look back at the results, the lessons learned and expectations for the future.

2. There are many reasons to study the internet as a data source for statistics. We mention a few of them. First of all, today the internet cannot be seen as just another communication channel: in many situations it is quickly becoming the *main*, and sometimes the *only*, communication channel. Shopping, travel arrangements, hotel and restaurant reservations, car maintenance scheduling, job vacancies, second-hand goods, in the Netherlands these are all increasingly done online. As the internet is increasingly becoming a normal part of everyday life, official statistics cannot afford to lag behind. Otherwise, they risk developing a bias towards traditional channels, a bias that may become larger every year. Therefore, it is important to start using internet data sources as well.

3. Secondly, using the internet as a data source may provide *new opportunities*. It can be a more efficient way to collect data that would otherwise be costly to collect or involve response burden. For example, many prices traditionally collected by telephone, written questionnaires or by visiting shops can be collected by internet robots via the web. Moreover, by making use of the huge amount of price and product information available on the internet, the number of observations and speed of processing may increase drastically, which could result in new fast price indicators. The same may apply to other statistical domains.

4. A third reason for studying internet data sources for statistics is that they allow statistics offices to observe, and maybe also understand, data patterns better than they can with traditional data collection. For example, traditionally the price of an airline ticket is collected manually several times a month. Using internet

robots, the price may be observed *more frequently*, daily, hourly, even every 10 minutes if you like. We have seen this in other statistical domains as well. Only by collecting data very frequently for a certain period, can we learn something about their volatility and thus decide what an optimal collection strategy would be. In a way, it feels like we can suddenly see in the dark, where before we could only touch and guess. This can put things in the right perspective.

5. However, all this does not mean it is easy. One thing we have learned is that it is good as well to distinguish between two quite different types of use: (i) collection of data from internet sites with many similar items, e.g. prices and characteristics of TV sets in consumer electronics web shops, and (ii) collection of data from internet data sources with only few items, e.g. the price of a cinema ticket from a cinema website. The first case, which we call *automated data collection*, can be approached with advanced internet robots that run without user interaction weekly or even daily to collect larger portions of observations at each run. For the latter category it would become too expensive to write dedicated robots for every single website and therefore quite a different approach was chosen. Here, we chose to assist the data collector with a tool to work more easily with internet sites and data. Hence we call this approach *robot-assisted data collection*.

6. In this paper we describe both types of use. Section II reports our activities with automatic data collection by internet robots for statistics. We describe the functionality of the generic base layer of software we use for most of the robots and touch on the infrastructure necessary to keep it running. We also look briefly at the legal aspects, and give some examples of characteristics of data generated by such robots. In section III, we dive deeper into robot-assisted data collection. We explain the concept of the so called “robot tool”, designed to help CPI specialists detect changes in prices on websites more easily. Lastly, section IV contains our conclusions from about five years of working with internet robots for official statistics.

II. Automated data collection from the internet

A. The concept explained

7. In short, internet robots are programs that imitate the behaviour of a human being visiting websites. There are many other terms for this kind of programs - crawlers, spiders, scrapers, bots - each with its own specific meaning, but we shall stick to the term internet robot throughout this paper. It is useful to note that internet robots are not new. They have existed just as long as the internet itself. In fact, many of today’s services rely heavily on automatic processing of internet information: there would not be any search engines, for example, if there were no bots crawling the internet. The general principle of an internet robot is sketched in Figure 1. Instead of a human operating a browser to request internet pages from a website, a robot program operates autonomously to issue the same requests to the website. It processes the information received and stores it in a database or data files. The robot, certainly a statistical robot, is usually not interested in graphical information such as style sheets and images, and usually does not even ask the website for them.

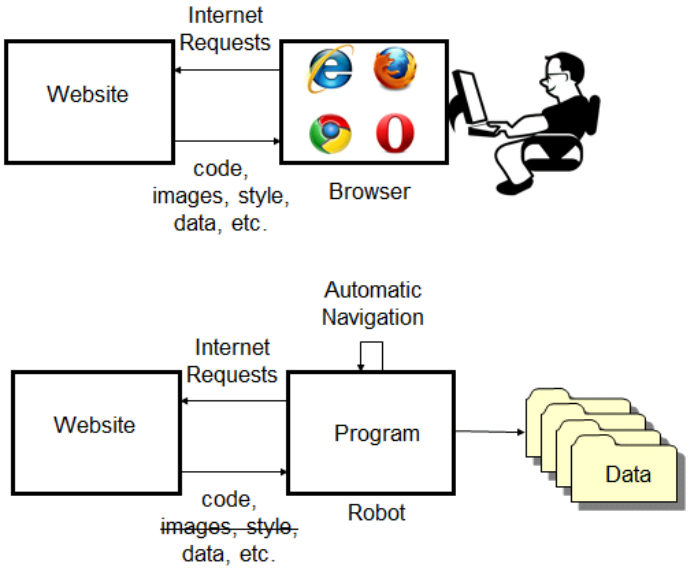

Figure 1: Human browsing compared with an internet robot

8. Technically speaking, internet robots are programs that issue http requests and interpret the response to identify and store relevant data. Over years, we have done quite a lot of experiments with different tools and technologies in this field. Simple versions of such robots may be implemented using user friendly tools with point and click functionality, that make it easy for non-programmers to automate repetitive tasks. More advanced tools on the market combine point and click functionality with programming extensions. In our view, point and click tools have limited applicability. We think automated data collection for official statistics cannot do without a fully featured programming language. In addition we found that some common functionality, required for almost any robot, could better be implemented in a generic base layer of software. This led to the development of the so-called robot framework described in section II.E.

B. Early experiments

9. One of the first organisations to use internet data for statistics was MIT [1]. They experimented with automatic collection of prices from online retailers and showed that it was feasible to calculate price statistics from internet data. Inspired by these results, Statistics Netherlands started experimenting with automated collection of price data from the internet for the consumer price statistics in 2009 and 2010 [2, 3]. Air ticket prices and fuel prices were collected daily by internet robots developed by Statistics Netherlands and by two external companies. These experiments showed that it was feasible to collect such data on a daily basis and the experiment ran for several months.

10. As a result the information we obtained on air ticket prices was much more condensed than the data from traditional collection. This made it possible to study such price changes in a quite different way. Figure 2 shows the results of the robot measuring the price of a ticket from Amsterdam to Milano (Malpensa), together with the results of the manual collection in that period. Although it was a bit difficult to compare the results (because of differences in definitions), the graph clearly shows a common trend. Both the robot and the manual collection observe lower prices in May and a substantial increase in June and July. Less prominent, but notable is that both show a small price drop during mid-March. From this graph we conclude that the manual collection actually approaches the actual prices quite well. Note that we can only conclude this because we measured the phenomenon more frequently for a limited period of time. From a more philosophical viewpoint one might say that we can suddenly see what could not be seen before. Of course, one should question the correctness of the robot observations as well. In order to get a feeling for that, we asked two of our robot partners to measure the same prices independently based on different robot technologies running at different locations for a few months. The results were almost exactly the same.

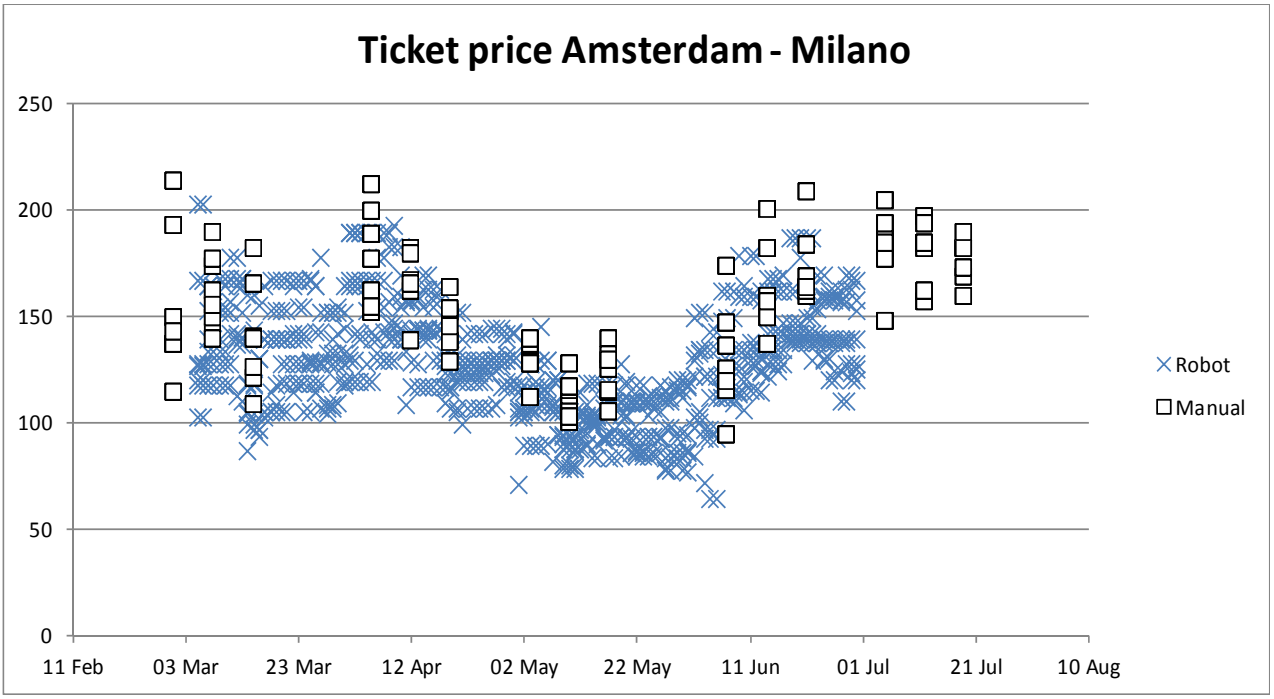

Figure 2: Ticket prices according to robot and manual collection

11. The above example shows how robots can be used to verify existing collection methods. But when we have a robot that is capable of obtaining data from the internet, we can also use it in very other ways than imitating existing collection. Figure 3 shows the result of a robot measuring the price of four airline tickets starting 116 days before departure up to the departure date (this figure was published earlier in [2]). This gives us a feeling for the volatility of these prices over time in a completely different way than for the CPI. We think this is typical for opportunities in this field: it is not only about doing the same things in a new way, it is probably even more fruitful to try to do things completely differently.

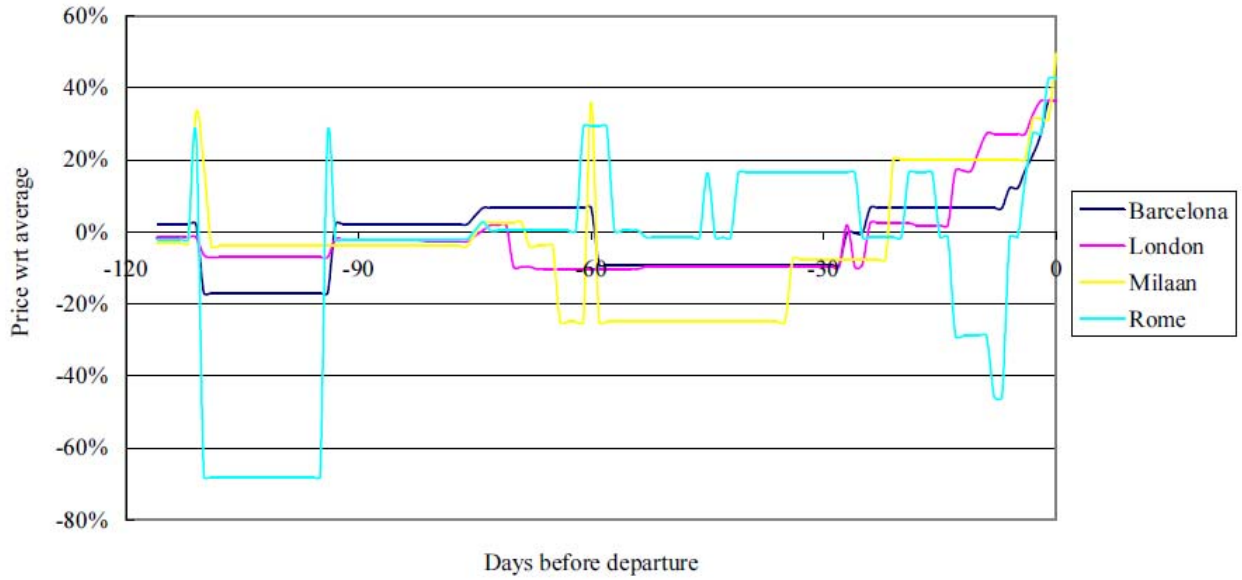

Figure 3: Volatility of flight prices of four destinations starting 116 days before departure.

12. Although these experiments were successful in a technical sense, there were doubts about cost efficiency, methodology and legal aspects. With respect to cost efficiency, we wondered whether manual collection would outweigh collection by robots, knowing that a robot would have to be maintained

technically for every small website change, which could become expensive. With respect to the methodology, we recognised that processing huge volumes of internet data would have its impact on the required statistical methodology. With respect to the legal framework, we concluded that further legal advice was required. We refer to [2] for a more detailed discussion of these issues. Despite these considerations we also felt that the only way ahead was to carry out further research.

C. Internet robots for the housing market

13. In 2011, we looked into the possibilities of generating fast new statistical indicators for the Dutch property market from internet data. At the start of the project many things were unclear. We did not know which sites would offer reliable data, which sites would offer original content and which would replicate or partly replicate the content of others, how easy it would be to read the data, which variables were available and how comparable they were across different sites. In addition, we did not know how the volume of data would grow and how volatile the data were. This is typical for internet data collection. Unlike more traditional data sources (administrative sources, questionnaires) where data characteristics are known by the delivering organisation, or controlled by the statistics office in case of questionnaires, statistics offices do not control internet data. It is more like observational data. We have to make do with what we can get.

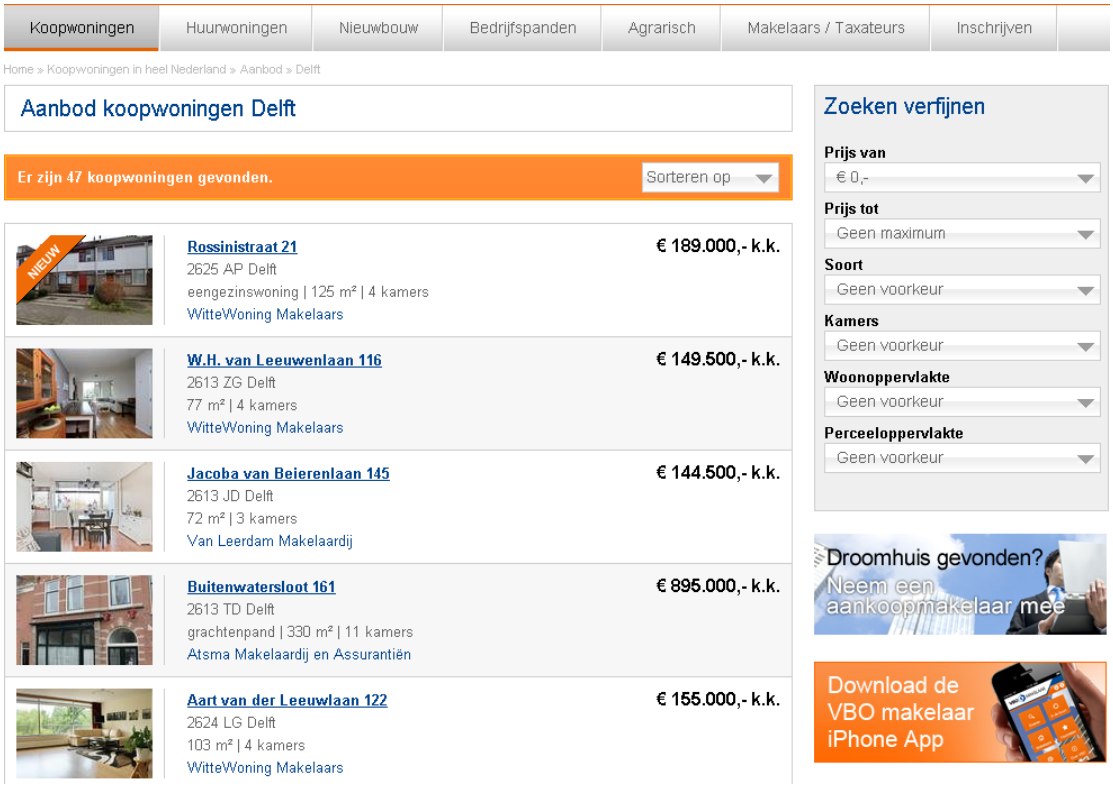

Figure 4: Example of a property site in the Netherlands

14. We examined about 30 Dutch property sites to identify characteristics such as the number of properties for sale, the variables available per property, the underling organisation, the stability of the site and some technical characteristics. Figure 4 gives an impression of the kind of sites we looked at. We made a distinction between sites with 'original content' and sites that collected data from others. This was not always immediately clear. Then we used internet robots to collect data from four sites for a small region in the Netherlands for about 12 months. As an example, Figure 5 below shows the number of new properties (positive bars) and the number of properties removed (negative bars) on one site in a period of about 190 days. It clearly shows a weekly pattern of activity, with fewer changes in weekends (but not 0). We could not easily explain the larger number of additions and removals on days 150 and 160, but we presume it was related to cleaning up the site's back-office system. We found this to be typical for internet data: there may

be unaccountable dynamics, and this should be taken into account when designing a robust methodology for processing these data.

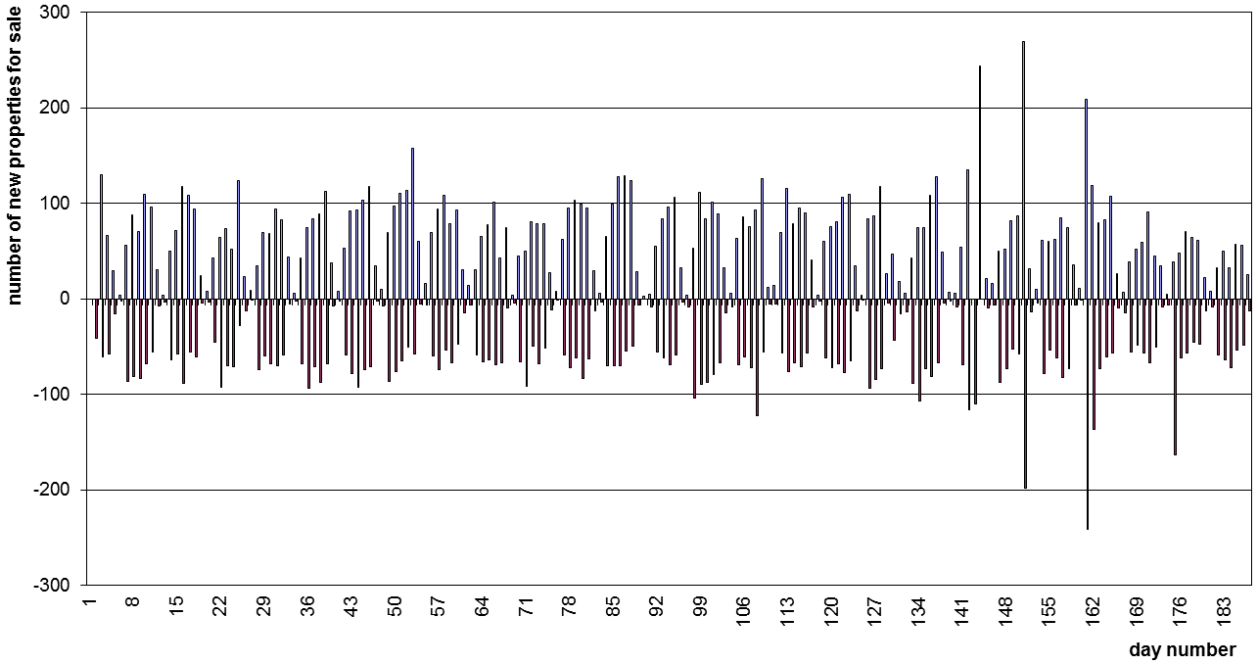

Figure 5: Volatility of the content of one of the sites

15. In addition, we identified the overlap in data between the sites, and compared the speed of changes in the content of the sites. For two similar sites we discovered that one of them detected 43 percent of additions and removals of properties faster. We concluded that this was the leading source of content. However the overlap analysis and speed comparison also indicated that at least one of the other sites was also relevant because it had content not covered by the leading source that could be statistically relevant. From a statistical viewpoint, these experiments can be seen as a way to increase our knowledge of the coverage of the sites with respect to the total statistical population. We compared online data with administrative data on property transactions (Dutch land registry) to get a feeling about the coverage of internet data with respect to the housing market as whole.

16. This exploratory data collection was useful to increase our knowledge about the data opportunities for property market statistics. Based on these results, Statistics Netherlands started negotiations with relevant organisations. At the time of writing this paper, Statistics Netherlands maintains two robots for reading data from property sites, with permission of the site owners. This results in a total data stream of about 200,000 records per week. These data are combined with those received from one other organisation to publish statistics on the state of the Dutch property market.

D. Clothes prices for the consumer price statistics

17. In a similar way to the activities for the property market, we experimented with automated data collection of prices of clothing for calculation of the Dutch CPI. We chose this domain because of the opportunities to reduce the number of shop visits: it is more efficient to download price information from web shops than to physically visit clothes shops, and it also reduces the burden for shop owners. And these days most shops also have a web shop.

18. So in 2011 we started to collect data from two major clothes retailers. Six other sites were analysed for their potential for statistics. Based on data collected so far, we can say that compared to the property market data there is one striking difference: these data are less structured. Each shop has its own way of naming items, and classifying them into menus and categories on their websites. This poses problems for

calculating statistics from multiple data streams from a large number of shops. One main challenge in this area therefore is to classify this kind of data automatically and to develop statistical methods to incorporate them automatically into the Dutch CPI. Research on this is still on-going, but the results are promising.

E. Robot framework and infrastructure

19. Although the data collected from the internet may vary drastically per statistical domain, we observed some general functionality that was required by almost all robots. Most web shops offer customers various ways to select goods, for example by category, price range, brand or any other specification. This usually results in a list of items being presented to the user, spread over a number of pages which can be navigated using previous and next buttons or page numbers. The overview page usually presents a number of general characteristics of the item, such as name, price (sometimes a sales price and an original price), a short description, etc. Usually, clicking on the items in the list generates more detailed information on that specific item. Figure 6 shows an example of an international web shop that fits this general pattern[1].

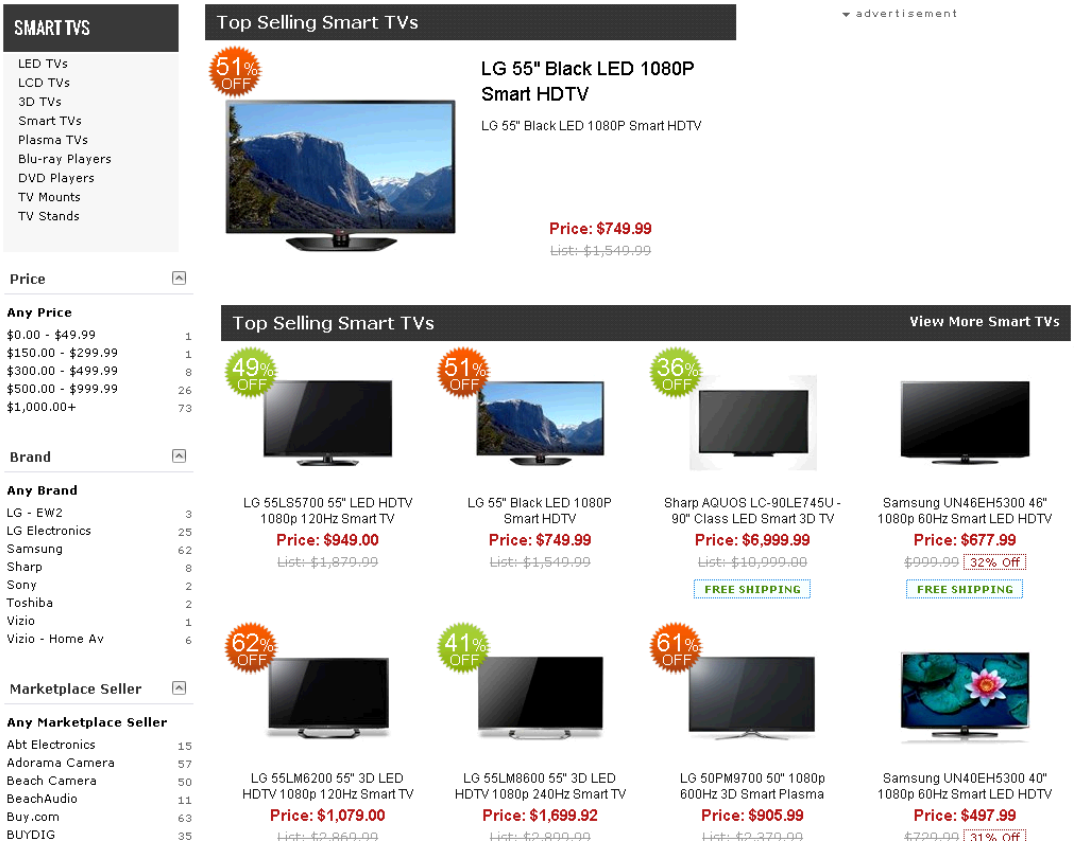

Figure 6: Example of a web shop that could be observed using the robot framework

20. The generic functionality described above is implemented in a so called *robot framework*. Sites that conform to the pattern described above only have to be configured via Xpath expressions specifying the navigation and the variables to be observed. For sites that do not fit the pattern completely, only the part that differs has to be added in a piece of code. All our robots are built on this generic framework. Technically, we implemented this base framework in the R programming language, but in principle any internet-capable language can be used. We chose R because it aligns well with the statistical data processing that follows the collection process.

[1] This is just an example; Statistics Netherlands did not collect any data from this web shop.

21. All robots run on virtual hardware connected to the internet. Their output is transferred automatically into Statistics Netherlands' production network daily. The size of the output data is monitored automatically, so that action can be taken in the case of failures (which may have many different reasons). We are currently considering adding an automatic restart script to all our robots, as this has proven to be successful in coping with robots that suffer from incomplete data collection.

F. Legal aspects

22. In [2] we already mentioned that legal aspects are important and need to be studied in more detail. A few years later, we feel that this field has still not been completely covered. In addition, regulations vary from country to country. Therefore, we took the practical approach to respect general applicable regulations and etiquette as much as possible, to be open and transparent about what we do, and to communicate with website owners whenever applicable and feasible.

23. Generally speaking the regulations and etiquettes we found out to be applicable are (a) internet etiquette (netiquette), (b) intellectual property rights (database law), and (c) privacy rules (national and international). With respect to netiquette (a), we respect the *robots exclusion protocol*, which may be used by sites to define paths that may and may not be visited. Also, in order not to influence site performance negatively, our robots are configured to run 'nicely', respecting a commonly accepted waiting time of one second between requests. To operate as transparently as possible, our robots identify themselves as being from Statistics Netherlands via the user-agent string. With respect to intellectual property rights (b), up to now we have always contacted site owners upon retrieving larger portions of their website. We do not bother them when performing minor tests to find out whether the content is potentially interesting for statistical use. With respect to privacy rules (c), most of the data we have retrieved so far cannot be related to individual persons. Where applicable, we issued a privacy statement about our activities, published on the website of Statistics Netherlands. In addition to the above general guidelines, whenever we have any doubts, we always contact the site owner.

G. Other issues

24. Once we started putting internet data collection into practice, as described in the above examples, we faced a number of unexpected issues. One of the things we have learnt is that although it is good to study the possibilities of internet sources for a while first, it is also very fruitful to start communicating with website owners. They know their data better than anyone else. Some offered to send us the data directly from their back-office system. If offered, we always opt for the direct connection rather than the robot solution, as it is expected to be more stable and may even contain more interesting data such as numbers of items sold (as in scanner data).

25. However, every situation has its pros and cons. Our experience so far indicates that in terms of time to market we can obtain internet data from robots much faster than we can obtain data from company back offices. So in some cases this might be a reason to choose for a flexible internet robot. In addition, it is useful to note that a direct connection with a back office needs to be maintained as well. In one case, our partner updated the website and forgot to update the connection from the back-office system to the statistics office, which resulted in a data gap of several months. If the data had been read from the site directly, the website owner would only have focussed on the change on the site and the statistics office would have been responsible for adapting the robot software – if necessary at all.

26. Another interesting case that came up when communicating with a site owner was that he also maintained an API (application programming interface) to give partners access to his data. He opened it for us and we now use it in combination with the generic robot framework to access the data. This would never have happened if we had not been in contact with the site owner.

III. Robot-assisted data collection from the internet

A. The concept explained

28. In 2012 we studied the possibilities of using internet robots to collect prices for the CPI formerly collected by phone. This would reduce the response burden for respondents, as prices would be read automatically from websites instead of by bothering people with a phone call. Two categories that appeared to be good candidates for this were prices of cinema tickets and prices of driving lessons. These categories were taken as examples to experiment with this new collection method.

29. In the cases we selected, it turned out that much of the information collected by phone was indeed available on the websites of cinemas and driving schools. However, previous studies [2] had shown that building a dedicated internet robot for every single respondent would be very expensive. In the cases chosen here this would have meant the creation of internet robots for 56 different sites, each with a different layout and structure. In addition, examination of past prices collected monthly from these respondents showed that volatility was quite low. These two observations led to the idea that it may be more profitable to build a tool to assist the CPI specialist to observe price changes on a site than to actually build an internet robot for each site. The idea was to build a tool to detect price changes on websites easily. In other words, the basic functionality of such a tool would be:

to speed up the work of the CPI specialist based on an automatic detection of changes in web pages.

An important additional requirement was that:

small website changes could be handled without the help of IT specialists.

B. Validation of the concept

30. Discussions with CPI specialists indicated that such a tool could seriously speed up their work. However, there were two important practical characteristics that would have to be taken into consideration:

1. When collecting prices from websites, the CPI specialist always starts on the homepage and follows the site menus to get to the required information. This is crucial to ensure that prices are not taken from parts of websites that are still available, but not actively maintained.

2. The prices to be collected may be part of a page where many more prices, ads and other information are available. If the tool was be able to detect changes on the whole page only, it would be less valuable. The CPI specialist should ideally be able to decide which parts of the web page to check.

31. Knowing this, we decided to look around for an existing tool with this functionality. A study on the functionality of existing tools and browser add-ons for website comparison in 2012 showed that existing tools and plugins did not offer the functionality required. More specifically they lacked functionality to check:

1. *the total path through a website*:
   As explained above, this is necessary to verify that the product concerned is still available via the site's homepage.

2. *specific parts of a webpage:*
   As explained above, this is necessary to narrow the scope of the automatic check to the exact context of the product concerned.

32. We subsequently developed a simple prototype to visualise the idea and enable the CPI specialists to test the concept. This we called the *robot tool*. This early prototype was used to turn it into a more professional version for collection of prices in 2013[2].

33. It is important to mention here that, in contrast to the use of autonomous internet robots for collecting data from websites, in our view the legal aspects mentioned earlier do not apply to the same extent here. After all, the robot tool only assists the end user in doing the job more easily than by hand and does not take over the process. In fact: is it a robot at all? We don't think it is, but it is debatable. Also, this tool is designed to check only the statistically relevant parts of a website for changes, and does not read large portions of the content. This differs from the approach described in the previous chapter.

C. From concept to tool

34. In short, the tool imitates the navigation of the CPI specialist to consult the price of a product on a website. It follows the path through the website the CPI specialist would otherwise have followed. The ability to check the complete "*navigation path"* through a website is important, as the CPI specialist can then verify that the price to be collected is indeed a valid price and not a price on an obsolete page.

35. Another important feature is that the tool enables the user to check specific parts of a web page instead of the whole page. If the tool compared web pages as a whole, it would detect many more changes, creating false alarms which would result in more work. By focusing on the specific part of the internet page that contains the required price information, the tool can be used more precisely and more effectively. We call this specific part of the last page in the path to be checked the "*price context*".

36. Having defined the *navigation path* and *price context* concepts, we can describe the basic functionality of the tool in more detail. This functionality is: to compare the price context with the result stored by the tool the last time it was executed. If the two results are identical, the price that was previously stored together with the previous price context is automatically saved together with the new price context. If they differ, the differences are displayed to the user. The user then has the options to copy the previous price (the change in the price context has no implication for the price), to enter a new price (the change in the price context does have an implication for the price), or to visit the website to study the change in the price context in more detail.

37. The user can organise the websites to be checked in so-called "Product groups". All websites of a product group can be checked sequentially without user interaction by clicking on the play button for that product group. The tool then checks all websites of that group and displays the result of the check using traffic light colours. Figure 7 gives shows the tool, configured to check the prices of some cinema tickets and the Ikea *Billy* bookcase[3] in various countries. It shows the first 10 of a list of 22 websites where this bookcase can be found. For most of the sites, no difference was found (prices in green boxes), but for one a difference was detected (-1 in the red box).

[2] Part of this work was done in the scope of a Grant project subsidised by Eurostat.
[3] This example was chosen because it is a well-known product in many countries. The prices shown are not included in the Dutch CPI.

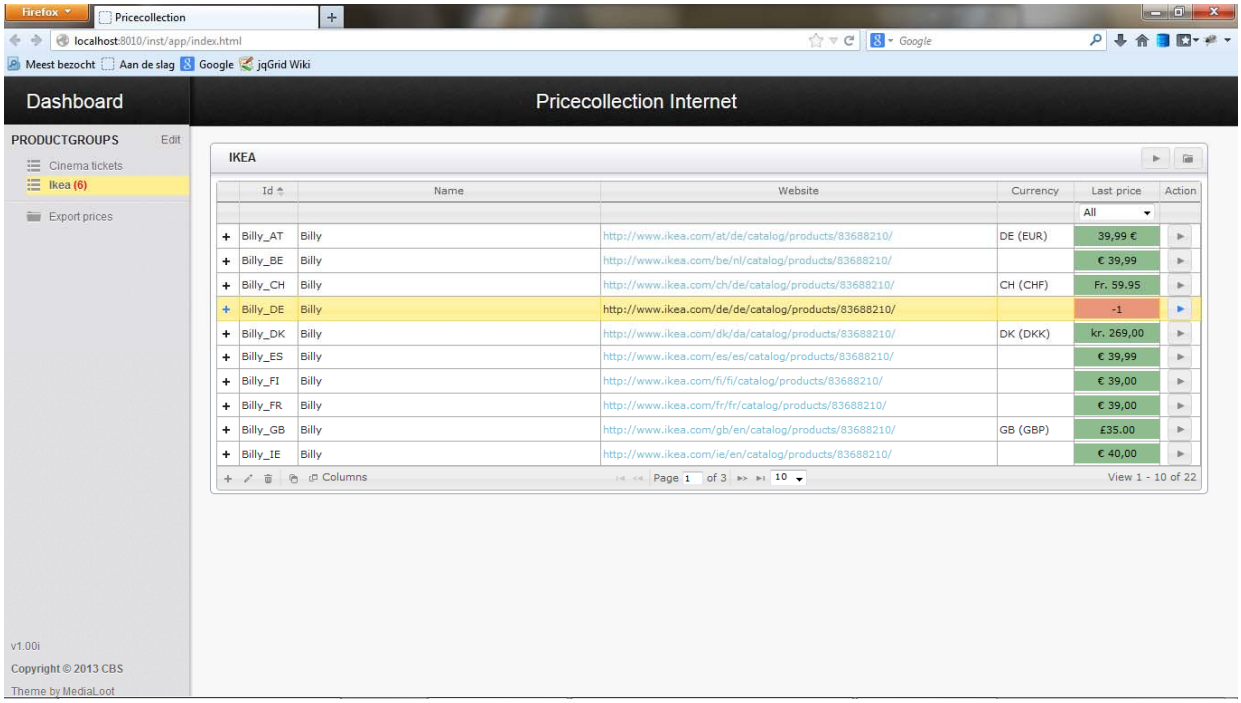

Figure 7: Robot tool

38. When the user selects the missing price, the tool displays the price context of the most recent run and the price context of an earlier run in a detail window. Figure 8 shows this detail window for the missing observation from Figure 7. Changes in the text are highlighted in red and green.

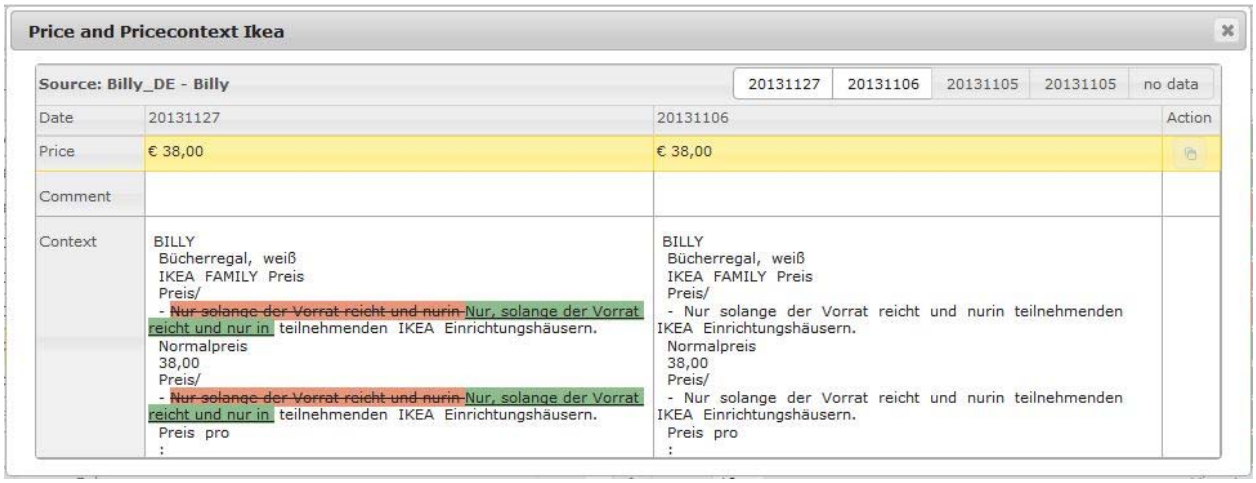

Figure 8: Detail view of robot tool

39. This detail window shows that an error in the text on the site has been corrected ("nurin" was changed to "nur in" twice). As the price was not changed, the user may conclude that the old price is still valid and, with one click, instruct the tool to store the old price in the observation database.

40. In order to instruct the tool which price context to compare over time, the CPI specialist has to configure the sites to be visited and the navigation path through the site. This is done with so-called *Xpath expressions*, a well-known internet technology. Further details about this technology would go beyond the scope of this paper.

41. The tool cannot be used effectively in all situations: prices expressed in pictures (jpg, png), in flash or in PDF cannot currently be compared. If this becomes a priority, further research could be carried out into what extent these media could be supported.

IV. Conclusions

42. This paper reports on our internet robot activities of the past few years. We have learned that it is useful to distinguish between (i) collection of data from internet sites with many similar items and (ii) collection of data from internet data sources with only a few items. The first case, which we call *automated data collection* can be approached with internet robots that run without user interaction. For the latter category, we developed a tool to assist the data collector to check for changes in data on internet sites. We call this approach *robot-assisted data collection*.

43. Both automated and robot-assisted data collection have proven to be viable options for official statistics. Automated data collection can result in more detailed data compared to data collected in traditional ways, which may be used to validate our work, to improve efficiency or to reduce response burden. Also, this kind of collection methods may be used to study phenomena in a completely new way. Robot assisted data collection appeared to be useful to collect prices from many different internet sites in an efficient way. A first version of the robot tool is taken in production in January this year. Using the feedback of the data collectors we plan to enhance the functionality of the tool further.

44. In general, if we try to look into the future we think a lot still has to be done before the use of internet data will be common practice in official statistics. One challenge is to process less stable and less structured data automatically into meaningful and reliable statistics. We plan to work on more intelligent internet robots to realise this. In our view these “second generation” robots will work differently, for example they would have the ability to automatically recognise data relevant for statistics. Also, instead of reading data per site, they should be able to combine data from multiple sites semantically and select the statistically relevant details, and maybe process them as well. Lastly, we have seen that this work has led to a number of spin-off projects, such as using search results to classify fuzzy data, automatically geo-coding data which were previously manually corrected, and using information from Wikipedia on business as an additional source of information for validating and improving the data we already have. We see that this type of technology can be used for other purposes than data collection. Of course, in a world where the internet does not end at the national border, this kind of research should be carried out in close international cooperation. We would therefore like to invite colleagues from other countries to join in.

V. Acknowledgements

45. The work described in this paper could not have taken place without the support, knowledge and advice of many colleagues, both at Statistics Netherlands and international colleagues, especially from Germany and Italy. At the risk of forgetting to mention some people, we want to thank our colleagues Karlijn Bakker, Dion Dieleman, Robert Griffioen, Nico Heerschap, Guido van den Heuvel, Rutger Hoekstra, Edwin de Jonge, Vincent Snijders, Peter Soons and Karine Tanis.

VI. References

[1] *Scraped Data and Sticky Prices*, Alberto Cavallo, MIT Sloan, December 28, 2010

[2] Hoekstra, R., O. ten Bosch and F. Harteveld, 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.

[3] *New data sources for statistics: Experiences at Statistics Netherlands*, P. Daas, M. Ros, C. de Blois, R. Hoekstra, O. ten Bosch en Yinyi Ma, NTTS conference, 22-24 Feb 2011, Brussels