**Statistics Netherlands**

Division of Macro-economic Statistics and Dissemination
Development and Support Departement

*P.O.Box 24500*
*2490 HA Den Haag*
*The Netherlands*

# Automated Data Collection from Web Sources for Official Statistics: First Experiences

**Rutger Hoekstra, Olav ten Bosch and Frank Harteveld**

Remarks:
Corresponding author: Rutger Hoekstra (r.hoekstra@cbs.nl)
The authors would like to thank Jelke Bethlehem, Piet Daas, Lisette Eggenkamp, Jan de Haan, Heleen Hanssen, Nico Heerschap, Els Hoogteijling, Mark Kroon, Douwe Kuurstra and Huib van de Stadt, Eric Wassink en Bert van Zanten and the members of Statistics Netherlands' Advisory Council for Methodology and Quality for their comments.
The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

*Project number:*
*BPA number:* *2010-132-KOO*
*Date:* *June 30 2010*

# AUTOMATED DATA COLLECTION FROM WEB SOURCES FOR OFFICIAL STATISTICS: FIRST EXPERIENCES

*Summary: As the internet grows National Statistical Institutes (NSI) are increasingly using data from the internet as a source for their statistics. For example, at Statistics Netherlands, the Consumer Price Index (CPI) department collects data from the web on articles such as airline fares, books, cd's and dvd's. The data is collected manually by statisticians that visit websites, locate the required data and copy it into the CPI-database.*

*This paper describes the first experiences of Statistics Netherlands in the automated collection of data from web pages, an approach which uses a technique popularly referred to as "internetrobots". This paper investigates two important research questions.*

*First, we explore the **feasibility** of this method of data collection for NSI's. We conclude that from a technical point of view this type of data collection is possible. Furthermore, the use of website data is probably legal, but we are not yet able to answer this issue with legal certainty. It is also concluded, that methodological applicability will have to be assessed on a case by case basis.*

*Second, we investigate the **desirability** of this type of data collection, i.e. do the benefits outweigh the costs? On the basis of six case studies (websites for four airlines, one housing site and one filling station website) tentative conclusions are drawn. It is difficult to gauge the cost-effectiveness of the robots when compared to the manual data collection process. Our tentative results do however show that the cost of reprogramming the robots in case of a site change and the initial investment in the manual process are the most important factors. Furthermore, we conclude that efficiency gains are likely to occur due to the learning-by-doing and scale effects. Finally, we conclude that the main advantages of internetrobots are not to replicate existing manual processes but to download more data (leading to increases in quality, frequency and speed) and to implement new areas for which manual data collection is not an option.*

*Keywords: internetrobot, webcrawler, data collection, web scarping, web spidering,*

## Table of contents

## 1. Introduction

The internet has evolved into an essential part of life in the developed world. It has become the primary way in which companies, governments, institutes and people communicate and do business. The internet is therefore becoming more and more important to National Statistical Institutes (NSI) because the economic and societal impacts of this domain have to be captured in statistics. However, the internet also provides another opportunity to NSIs. During this communication process a plethora of electronic information is published on web pages on the World Wide Web. In some cases this data can be used by NSIs in the production of official statistics.

The process of using the web as a data source is already underway. For example, the consumer prices index (CPI) of Statistics Netherlands is partially based on data from the internet. Statisticians surf to specific web pages to extract prices for various CPI-articles such as books, airline tickets and electronics. Special protocols have been developed per CPI-item to ensure that statisticians collect the data using a consistent and structured statistical procedure.

It is likely that the use of the internet data by NSI's will increase in future. There are a number of trends that make this probable:

1. *Expanding information base*. The amount of data that companies, institutes and governments publish on the internet is increasing all the time. It seems likely that this will lead to an increase in the amount of statistics that could potentially be created using websources.
2. *Aggregator sites.* There is a large increase in the number of websites that collect and bundle information from other sites. For example, there are many sites that compare the prices of electronic goods from different suppliers. From a statistician's point of view this means that a larger share of certain field is covered by a single website in a harmonized format. In the Netherlands, for example, websites for the housing market are competing by offering as many dwellings as possible -from different sources- on their website. It must be noted that from a methodological point of view this type of site has the disadvantage that there is no guarantee that the content of the aggregator site is equivalent to that of the source page. Also there may be a selection bias in the underlying web pages which the aggregator site covers.
3. *Increasing relevance*. The importance of the internet in our daily lives also means that the demand for statistical information about the internet is also becoming more relevant.[1] A couple of years ago the internet provided a

[1] The increasing importance has encouraged Statistics Netherlands to start a program for statistics about the "Impact of ICT on society". The program will develop new statistics on the economics and societal impact of information and communication technology but also includes innovative data collection methods. Starting in 2010 the project which is described in this report will be part of this statistical program.

(cheap) means for CPI-statisticians to find a proxy of prices in the actual stores. However this situation has changed because many markets (e.g. holidays and books) are quickly going online. The online prices that are collected are therefore no longer proxies, but the actual object of study for a large share of the market.

The statisticians at Statistics Netherlands that collect internet data do this manually. They visit the relevant website, collect the price data and transfer the data to the CPI database. However, websources are obviously electronic and therefore raise the prospect of an automated data collection process.

There are several ways of collecting data automatically using computer programming techniques. We will refer to all these technique using the term "internetrobots" but they are sometimes also referred to as "webscrapers", "webcrawlers" or "webspiders". The simplest way to explain an internetrobot is to think of the manual process: a person visits a site, collects the data and transfers it to a database for statistical use. This is precisely what an internetrobot does as well. However, since it is a computer rather than a person it means that the data collection period is much shorter and that the computer can be programmed to repeat this procedure every year, month, day, hour or minute, day or night. On the other hand, the internetrobot lacks the ability to check the validity of the website in real-time quite the same way as a human observer can.

The core research questions addressed in this report concern the feasibility and desirability of automated data collection of web sources:

1) Is it *feasible* for NSIs to adopt automated data collection from web sources? In this research question technical, legal and methodological preconditions need to be met.

2) Is it *desirable* for NSIs to adopt automated data collection from web sources? Here the costs and benefits have to be weighed.

To our knowledge no NSI has yet implemented an automated data collection process.[2] In this report we wish to start the thought process amongst NSI's about this type of data collection and to share the first experiences of Statistics Netherlands in this field. Specifically we focus on six case studies which we have investigated during the period June 2009-March 2010 (four websites for airlines (which are currently part of the CPI manual data collection process), one housing site and one site for an unmanned petrol station). Based on these case studies we also make

[2] However, Statistics Netherlands has previously collaborated in a project of the Ministry of Economic Affairs. The report, which was written by a consultancy firm, took a broad look at the use of electronic data for statistics, of which internetrobots are a small part (Dialogic, 2008). Previously, the methodological division of Statistics Netherlands has also suggested that the use of automated data collection from websites was a promising area for innovation in statistics (Daas and Beukenhorst, 2008; Roos et al., 2009).

tentative comparisons of the costs of the current manual data collection processes and the automated processes.

The structure of the report is as follows. Section 2 deals with the feasibility of automated data collection by discussing the legal, methodological and technical preconditions. In Section 3 we delve further into the benefits and costs that provide a basis for deciding whether this data collection process is desirable for NSI's. In Section 4, the experiences from six case studies are discussed. Specifically we focus on what technologies and which cases were chosen and what we learned in the process. Tentative cost efficiency calculations are also shown. In section 5 conclusions are drawn and the future work of our project are discussed.

## 2. Feasibility: Legal, technical and methodological preconditions

To assess whether this type of data collection is feasible for NSIs a number of preconditions have to be met.

First and foremost, this type of data collection has to be *legal*. It goes without saying that an NSI cannot adopt techniques which are against the law. Of course we have only been able to investigate the Dutch legal setting. The legal department of Statistics Netherlands point to two laws which are important in the Dutch situation: the "database-law" (Databankenwet) which regulates the protection of the investments that companies make in large databases; and the "CBS-law" (CBS-wet) which regulates the rights that Statistics Netherlands has in the collection of data.

On the basis of these laws a distinction can be made between "substantial" and "non-substantial" use of data. In both cases it is wise to inform the company. This is a matter of *netiquette* and good customer relations.[3] It is unclear, from a legal perspective, what happens if a company has objections to the use of their data. More legal advice is required to conclusively resolve this issue, particularly in the case of substantial use of databases, where the benefits of internetrobots are greatest.

The second precondition is that this type of data collection is *technically feasible*. To our knowledge, robot techniques have never been applied in the statistical data collection process but are used on a very large scale on the internet. The technical feasibility for NSI's therefore seems likely and is confirmed by the successful implementation of six case studies (see section 4).

A third precondition for an NSI is that this type of data collection is *methodologically sound*. As with any new data collection method, an NSI needs to know that the measurement strategy adheres to the standards of official statistics. How often is the information of the website refreshed? Is the online price different to the price in the stores? Are different webbrowsers displaying the same prices?

[3] Similarly it is a matter of good manners that the server of the targeted site is not overburdened by the CBS-robot. There are technical solutions to make sure that the CBS does not overburden external websites.

These are some of the more obvious questions when it comes to the applicability of webdata to statistics.

In this project we have been unable to delve into methodological issues due to time constraints. Additionally, most of the case studies investigate the replacement of existing CPI manual data collection, which means that the use of data from specific websites has already been found to be acceptable. Of course, the new method of data collection poses new methodological issues but given our resources we felt that the other two preconditions, technical feasibility and legality, were more important priorities. Furthermore, many of the important of certain methodological problems can only be investigated through 'learning by doing' (for example, how often does a site change?). Therefore, despite the fact that this was not the focus of our efforts many very relevant methodological aspects are discussed in this report.

As the internetrobots are applied to new areas, the methodological issues will become far more important. It is clear that websites will have to be assessed on a case by case basis, because it is very hard to make generic comments about the methodological issue surrounding websites. When assessing the individual cases it is good to know that the methodology department of Statistics Netherlands has recently produced a methodological checklist for future applications of internetrobots (Ossen et al, 2010).

## 3. Desirability: Benefits and costs

Even if the production of internetrobots is feasible it may still be rational for an NSI to use alternative data collection methods. The assessment of the desirability is simply a matter of weighing the benefits and costs of each strategy. That is why, in this paragraph, we take a closer look at the potential advantages and disadvantages. It is important to realise that there are four ways in which internetrobots can be applied:

A. Replace existing manual data collection

B. Improve existing manual data collection (speed, frequency, quality)

C. Replace existing survey-based or register-based collection techniques

D. New applications

When discussing the potential advantages we will indicate in brackets which of these types are affected. The potential advantages are:

- Efficiency (A, C). This is of course one of the primary reasons that NSIs may consider this type of data collection. The manual process is labour intensive and it is conceivable that an automated method would be more efficient. Because of the importance of this point we analyse it in more detail in section 4.

- Survey burden (C). If the internetrobot replaces an existing survey, it will reduce the survey burden on society.
- Timeliness (A, B, C). Because data is collected in real time, it is conceivable that this type of data collection will enable NSI's to publish faster than previously.[4]
- Frequency (A, B, C). An internetrobot can be programmed to collect data per year, day, hour or minute, at a fraction of the time that it takes the manual data collection process. This allows for more frequent observations which some fields of statistics demand.
- More extensive (B, C, D). An internetrobot can be programmed to enter all possible combinations of variables available on a website or to download all items from a huge amount of information spread across a big website. The amount of observations done by a robot exceeds what could be done manually, yielding more extensive datasets on certain subjects. This may also help to improve the quality of existing manual data collection. For example, in the case of data on air fares, Statistics Netherlands collects information from five carriers to five destinations. Given the resources of the CPI, it is understandable that they have limited the scope to this dataset. However, an internetrobot can, fairly easily, extract the information of *all* destinations of an airline. This means that the quality of the CPI statistics can be improved, assuming that the relevant volume data can also be found.
- Quality (A,C). In the previous bullet we discussed the improvement of quality by increasing the data input into the CPI methods.
- Hedonic price measurement (B). One of the advantages of internetrobots is that they can collect data about the characteristics of products. For example, housing sites have information on the asking prices of the houses but also on the number of rooms, plot size, garden etc. This is information which may be used in the calculation of hedonic price indexes. These are superior to price indexes that do not take other quality characteristics into account.
- We do not discuss work-related aspects which may also turn out to be beneficial to the worker. For example, the internetrobot may help to reduce RSI and lead to less monotonous work.

There are also a number of disadvantages or uncertainties that may occur:

- Changing web sites. One of the most important problems for internetrobots is a change in the website. In large scale restructuring of

[4] This also raises the prospect of producing indicators which are known as "beta-indicators" in the Dutch setting. These are indicators that are not as methodologically sound as existing ones, but which can be published more quickly. This is very useful if one wants information about the state of the economy while not having the time to wait for the official GDP figures.

websites, it is almost inevitable that the robot will no longer function and will have to be (partially) reprogrammed.

- Lack of human checks. In the manual process, a human observer is capable of checking the data during the collection process. In the case of a robot this is only possible by adding special modules into the program, but even then it is a matter of "what you put in, comes out". Subtle changes to websites, such as replacing a price inclusive of VAT to a price without VAT are only detectable if the robot has been given the task of checking for such anomalies.
- Resources. An NSI will very likely have to make investments in computers and personnel because programming internetrobots requires skills and software that are presumably not part of current core competences.

## 4. Case studies

### 4.1 Choice of case studies

When looking for valuable case studies for our internetrobots project we spoke to many people that had experience with web-based data. People tend to list a number of sites but these varied considerably in stability, quality, and usefulness for official statistics. Therefore we focussed on websources of which the information is currently used for the Dutch CPI. We also talked to the CPI-experts about potential new internet sources.[5]

Also, there is a project ongoing at Statistics Netherlands, which aims to improve the availability of data for the housing market. This project already has a lot of data from the Dutch land register and one of the largest realtor cooperatives (NVM). If online information was available it would become possible to analyse the asking price (collected by an internetrobot) and the selling price (from the land register). This may provide a valuable indicator of the tension on the housing market.

[5] Suggestions included were a) webpage's of airline carriers to collect airline fares to various destinations; b) webpage's of unmanned petrol stations to collect fuel prices; c) on-line catalogue of a mail-order firm (such as Wehkamp); d) on-line catalogue of companies with a large market share (such as Ikea); e) analysing and interpreting the information on the webpage's of a large Dutch websites that compares the performance and prices of electronics; f) webpage's of on-line clothing stores; g) compare on-line prices of new and used-cars; h) collect information of so-called "aggregator sites" that compare and present the functionality and price of many goods; i) automatic interpretation of the content of the Dutch Government Gazette (Staatscourant); j) collect price information of mobile phone subscriptions; k) collect prices of dvd's, cd's, and games; l) collect information of so-called "aggregator sites" that compare and present the functionality and price of many goods; m) use product specifications of technology vendors for price analysis.

Because of the combination of (potential) efficiency gains and the clarity of the cases it was decided to start to collect data from the website of an unmanned petrol station (www.tinq.nl) and of four airline websites (www.klm.nl; www.transavia.nl; www.easyjet.nl; www.ryanair.com). In addition we did some small tests with collecting prices of houses (www.funda.nl). The collection of airline fares are examples of internetrobots that could replace existing manual data collection ["category A" using the categories adopted at the beginning of section 3] while the use of the information on websites of unmanned petrol stations and the housing market are new fields of application [category D].

### 4.2 Choice of technology

From the very beginning of the World Wide Web many different types of robot scripts have been used. Search engines are examples of robots that have been around since the 1990s. These are proven technologies which have been developed an implemented on a large scale by Google and other search engine companies. At the same time, it must be said that the robot techniques are still in their infancy when it come to small scale use. We found only 3 companies in the Netherlands offered robot-services to other companies. This means that at this level the market has not yet matured and may be expected to change quite significantly over the next 10 years or so.

Obviously, when searching for the best technology various ways to carry out the internet aware tasks were compared. The use of open source *scripting languages* like *Perl* and *Python* are reasonable in this respect. These languages have strong internet facilities and there is a young and internet minded community that expands their functionality all the time. However, utilizing these languages requires specific programming skills which are currently not part of the regular capabilities of the IT (or other) departments of Statistics Netherlands.

Another approach is to use *search engines*, with their proven and powerful indexing mechanisms, to extract data from a specific website. Although the search engine suppliers we talked to guarantee that this is possible, it seems that these tools have a different focus. Search engines are oriented towards *indexing* and *searching* a huge amount of unstructured information. It is possible to augment the search engines with scripts that perform like robots, but this is an add-on functionality, and not the primary strength of these products.

A third way to go is to use dedicated robot tools. These tools are specifically developed to extract data from internet pages, sometimes by just indicating the information target visually and some additional scripting. The functionality varies from "point and click" methods (similar to "macro's" in Excel) to web scripting which is focussed entirely on robot functionality and is therefore more user friendly than the computer languages which were discussed earlier. In our short inventory we

found that there are already quite a few of these tools[6], which vary greatly in their technique and functionality. Tools require less programming skills, but they are less flexible than that of generic scripting languages. These tools are also more likely to be 'closed source' which suggest a better support of the user in comparison to the open source scripting languages. However, on the other hand the dependency on the company that provides the tool also increases.

The final option for NSIs is to *outsource* this task to a company that has in-house experience in this very specific field. Not only could the development of the robots be outsourced, but one might also ask them to operate the robots and deliver the data that is to be collected. There are a number of companies which specialise in these types of web data extraction.[7] We had some preliminary discussions with the three companies that we found for the Dutch market. It may come as no surprise that they felt confident that they can extract data from almost all websites and that they can do this very efficiently because of their experience in this field. After a bidding process we have selected two of the three companies to produce internetrobots for Statistics Netherlands in a pilot project.

In table 1 we have summarized some of the advantages and disadvantages of each variety of internet data collection.

*Table 1. The advantages and disadvantages of several varieties of automated data collection*

| **Internal/ External** | **Variety** | **Advantage** | **Disadvantage** |
|---|---|---|---|
| Internal | Scripting language | Programming with most flexibility<br>Usually open source languages | High-level programming skills as NSI |
| | Search Engine/Tool | Only moderate programming skills required at NSI | Programming not as flexible as computer language<br>Usually closed source language |
| External | As a service | No special programming skills required at NSI | Coordination with external parties may be problematic<br>Least flexibility when it comes to development of new robots or changing of existing ones |

It is not obvious which approach is most promising for the development of internetrobots at NSIs. Therefore, we decided to experiment with all approaches except search engines. Search engines seem to have a very different point-of-

[6] An inventory in the summer of 2009 resulted in the following examples: Irobots, OpenKapow, Newprosoft, Websundew, Iopus, Bget, Metaseeker, Happyharvester, Djuggler, Lixto and Screen-scraper

[7] An inventory in the summer of 2009 resulted in the following companies: Mozenda (US), Xtractly (US) and Scrapegoat (US). We also found 3 Dutch companies that provide robot services.

departure. They are designed to roam the web, by following links and indexing the web pages they find. Of course, search engines can be tailored to search a single page of website but this is not what a search engine is designed for. Given this observation, and the limited resources of our project, it seemed more fruitful to start with the other approaches.

Table 2 shows which case studies have been performed with the different varieties of robots. We have developed a number of pilot robots in Perl[8], a few in one of the best robot tools we found: Djuggler. As mentioned earlier, we also decided to have a pilot project where two companies develop, host and execute some robots for us for a period of three months. In this report we will only report on the results for Perl and Djuggler robots. The external parties which we have contracted will be running the scripts they developed in the period April 1st- July 1st 2010. Where possible we will report qualitative experiences from our discussion with the external parties but definitive results will be published in a subsequent report (end of the summer of 2010).

*Table 2. Case studies and technology*

| | | **Website complexity** | **Internetrobot variety** | | | |
|---|---|---|---|---|---|---|
| **Topic** | **Site** | | **Scripting language** | **Robot Tool** | **Company 1** | **Company 2** |
| | | | **Perl** | **Djuggler** | | |
| **Airline tickets** | klm.nl | Difficult | X | X | X | X |
| | transavia.com | Medium | X | | | |
| | easyjet.com | Easy | X | | X | |
| | ryanair.com | Difficult | p | | X | |
| **Housing market** | funda.nl | Easy | p | p | | |
| **Unmanned petrol stations** | tinq.nl | Easy | X | | | |

X – full implementation, p – partial implementation

There is one technical aspect which is extremely important for creating internetrobots: the communication patterns adopted by the website under investigation. In the 1990's, when the use of internet really took off, webpages were nearly all static pages coded in HTML (HyperText Markup Language). Although this is still in use, many sites use more advanced interaction patterns nowadays. Often, a simple response page based on some variables entered by the user, involves several communication requests between browser and a server loaded with databases. An internetrobot has to emulate these communication patterns to interact with the site in a way that reflects the behaviour of the user. Therefore, one of the

[8] The choice of Perl versus python was a practical one. Both languages have dedicated packages for web parsing, form-filling and dom handling. At this stage we cannot conclude which of the two languages performs better.

most important aspects of automation is analysing the way in which the website interacts with the user and the data.

The reason why a webpage is complex varies according to specific technologies. Examples are the complexity of form handling, the use of iframes, the use of session id's and cookies, the use of javascript, the use of ajax communication etc. Also, the technical design and organisation of the website, such as the way style and content are implemented, influences the time required to build the robot. In table 2 we have included a rough categorisation of the complexity of the interaction with the websites.

### 4.3 Experiences

*Existing manual data collection*

As part of the current manual data collecting process for the consumer price index of Statistics Netherlands, statisticians visually inspect the prices of several well defined airline flights. Ticket prices of the cheapest flights are chosen on a specific departure and arrival date one or two months ahead, for 5 destinations and for 5 different airline companies. This is done 3 times a month. With an internetrobot one could easily increase the number of observations (each day / multiple times a day), the number of destinations, and the flights being inspected (one week ahead, one month ahead, two months ahead, three months ahead etc.). However, we have chosen to start simple and emulate -as much as possible- the existing manual process with one exception: we executed the robots daily.

One of the first difficulties we observed when experimenting with the robots is the problem that, although the robot may perform correctly most of the time, it may also fail unexpectedly. This results in missing values. Further investigation showed that this was mainly the result of slow connections which resulted in time-outs. After experimenting with different time out values for internet requests, this problem was solved. After that the robots performed almost flawlessly.

One challenge when using internetrobots is to keep them working when web sites change. These changes vary from simple layout and style changes to a complete makeover in terms of technology and communication structure. Obviously, the volatility of the site and type of changes is an important aspect of a robot observation strategy. In table 3 we have summarized our experiences with respect to site changes and the time it took to reprogram the robot.

KLM appeared to have one major site change but its website underwent many cosmetic changes which had no effect on the functioning of the robot. The robot created with the robot tool worked for over 6 months without any problem before a site change made it stop working. The robot in Perl experienced the same problem at that point in time. The actual fix took no more than 8 hours. The robot for Transavia however worked for 3 months without any problems, until it stopped because of a site change. The fix took much longer, due to the fact that we also redesigned the

robot to operate in a more robust way. The amount of time spent on the fix and redesign was about 40 hours. The Easyjet robot has been operational since December 2008. It ran for over a year without problems. When the site did change it took about 24 hours to reprogram. The robot for Ryanair has still not been fully implemented.

*Table 3. Web site changes*

| | **How long is the robot operational?** | **How often did site change?** | **How long to reprogram?** |
|---|---|---|---|
| **klm.nl** | 5 months (Perl) | Once (February 2010) | 8 hours (Perl) |
| | 6 months (Djuggler) | Once (February 2010) | Fix not attempted. |
| **transavia.com** | 13 months | Once (June 2009) | 40 hours |
| **easyjet.com** | 16 months | Twice (January 2010 and March 2010) | 24 hours (January 2010)<br>March: Fix not attempted. |

The flexibility of a robot means that for a specific airline website, one can easily expand the range of departure and arrival characteristics of flights to be observed (or economy and business class for example). As an experiment we decided to create a robot for one of the airline companies that collected the prices of four specific flights in December (departure from Amsterdam on December 14th 2009 to either Barcelona, London, Milan or Rome, and return on the 17th). We started tracking these flights on the 18th of August 2009. This resulted in a dataset with daily observations of the prices for the 4 destinations starting 116 days ahead of the day of departure. The results are shown in figure 1. The prices are compared to the 116 day average fare.

The results show that booking a flight on the last day costs you 40% more than the average price of the flight in the 4 months before. This premium is fairly consistent for all destinations. The results also show that flight prices can fluctuate tremendously up to 70% below the average price (for Rome, about 100 days before departure) to about 30-40% above average (for Rome and Milan about 60 days before departure).[9]

[9] It is important to realise that the volatility of the price is caused by an automated interactive price strategy on the part of the airlines. Based on detailed information on the demand for tickets and the passengers, air fares are set in such a way to optimize revenue. In a sense there is a robot pitched against our robot. This field is called "revenue management" and even has its own scientific niche. It does however raise the question how much "noise" is in our data collection and which data we are feeding into their system with our robot. We are grateful for Prof. Franses of CBS's Advisory Council for Methodology and Quality for this comment.

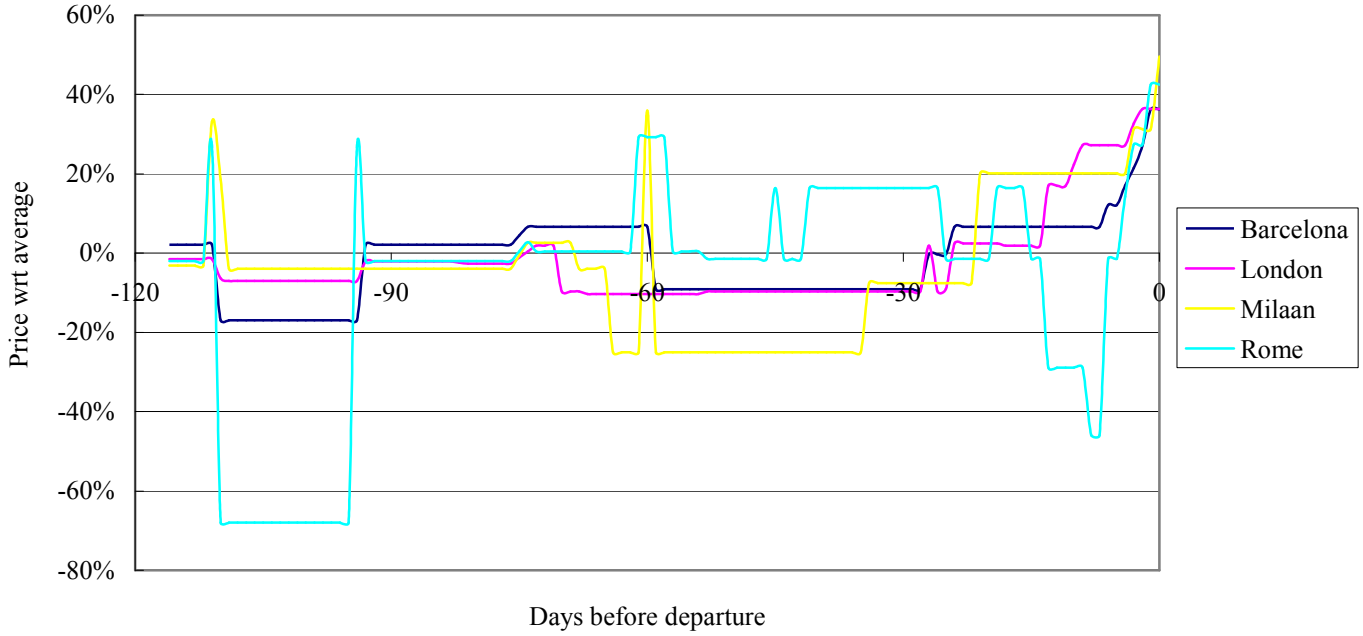


*Figure 1.Flight prices of four destinations starting 116 days before departure*

Now recall that the current CPI-method is to book a flight with a departure date one ahead. Figure 1 shows the ticket prices for Rome being sold at about 20% above average at that point in time. Assuming that this would not be the case the next month one would get a rather unrepresentative price index. Nevertheless the volatility of the graph also begs the methodological question how to calculate the "real" price of this flight. We will not solve this problem here, but simply suggest that this type of problems could benefit from the methodological advancements in the scannerdata index theory which also involves very volatile and frequent data. Nevertheless, it should be clear that this methodological question can only be investigated using data from internetrobots because the manual data collection would take to much effort and requesting the data from the airline companies would contribute to an increase in the survey burden.

With respect to the knowledge needed to create robots, we already mentioned above that understanding website communication structures is key to successful robot design. During our tests we found that especially knowledge of (Ajax) communication, the document object model, cookies and server variables is inevitable to make a working robot. In addition, knowledge of javascript is very important.

First experiences with the robots created by external companies show that their robots perform very well in terms of reliability and robustness. After two weeks of pilot operation there were no missing values. Also the robot for Ryanair proved to be no problem for them, while we were unable to implement it. As a result of our

exchange of experiences with these companies, we have a strong feeling that all information, except information that is embedded in flash sites[10], can be retrieved. However, in both modes of operation, do-it-yourself versus outsourcing, it will cost money. For that reason in the next chapter we try to draw some conclusions on the possible efficiency gains resulting from our robot case studies in relation to the costs involved.

*New applications*

The airline tickets robots are aimed at reproducing the existing manual data collection process. However, we also looked at two new fields in which robots may be applied.

The robot developed for the collection of product price information of an unmanned petrol station (www.tinq.nl) is the most successful workhorse in our pilot studies. This robot started collecting data at the end of 2008 and kept working up to March 2010 without any problems. Everyday it collected (and is still collecting) data on 4 types of fuels for over 200 unmanned petrol stations in the Netherlands.

We also did some small experiments with data collection from on of the most important housing site in the Netherlands (www.funda.nl). The data collected had some blanks and strange items, but the results indicated that it should be fairly easy to automatically extract data from this site.

### 4.4 Tentative calculations of cost-efficiency

One of the advantages that automated data collection may provide is that it is more efficient than the manual data collection process. In this section we provide some tentative calculations in which we compare the manual and automated processes. Note that the results are very provisional and are only meant as rough indication.

Table 4 shows the time spent on the manual collection of prices for airline fares. A statistician spends nearly two and half hours a month per carrier. The process takes little longer for klm.nl because of the number of destinations which are recorded. In the second to last column we have recorded how many hours will have been spent in the manual process after 3 years.

Now let us look at the results for the automated approach. Here we make the distinction between several types of costs: production costs; development costs and costs for the maintenance of the robots.

Production costs are included because the process will not be automated 100%. Statisticians will still need to do manual checks and upload the data into the CPI database. The development costs are equivalent to the programming time which was

[10] There are however efforts to improve the ability to search in Flash (see http://www.adobe.com/devnet/flashplayer/articles/swf_searchability.html)

required to produce the robot. This can be seen as the start-up costs of the data collection. Finally there are also the costs of the maintenance of the robot. The most important factor here is the number of times that the website changes and the time it takes to re-program the robot. As table 3 shows it is very difficult to distil a "rule of thumb" for this issue. We assume that each website changes once a year and that it will take 24 hours to adjust the robot to the new situation (which boils down to 2 hours a month).

In the last two columns of table 4 insights are gained into the efficiency of the internetrobot. As can be seen the break-even point for klm.nl is reached within 4 years (both for robots programmed in Perl and in Djuggler). This means that a sizeable chunk of CPI-work could be replaced very cost-efficiently. However, the "smaller" manual tasks perform quite poorly with no break-even scores above 20 years. This is actually an underestimation because we have not been able to realise a full implementation for the Ryanair website.

*Table 4. Break-even calculations for internal automated data collection (hours)*

| | **Total (hours)** | | | **After 3 years** | **Break-even** | |
|---|---|---|---|---|---|---|
| | **Production** | **Development** | **Maintenance** | | | |
| | Per month | One-off | Per month | Total | Months | Years |
| **Manual data collection** | 11,7 | 0 | 0 | 420 | | |
| klm.nl | 4,7 | 0 | 0 | 168 | | |
| easyjet.com | 2,3 | 0 | 0 | 84 | | |
| transavia.com | 2,3 | 0 | 0 | 84 | | |
| ryanair.com | 2,3 | 0 | 0 | 84 | | |
| **Automated data collection** | | | | | | |
| klm.nl (perl) | 0,7 | 96 | 2 | 192 | 48 | 4,0 |
| klm.nl (djuggler) | 0,7 | 81 | 2 | 165 | 41 | 3,4 |
| easyjet.com | 0,3 | 59 | 2 | 143 | >240 | >20 |
| transavia.com | 0,3 | 123 | 2 | 207 | >240 | >20 |
| ryanair.com | 0,3 | 263 | 2 | 347 | >240 | >20 |

The results in table 4 vary tremendously and are very much influenced by the crude assumption about the costs of maintenance. Furthermore it must be said that these are the first internetrobots produced at Statistics Netherlands. Although they were created by experienced programmers, they did not have specific experience in this field. It is therefore likely that there will be a certain learning effect as the portfolio expands. Code can be recycled and debugging experiences from previous cases can be reused. Also the scale with which this type of data collection is implemented at Statistics Netherlands is important since it is likely that as the number of robots increases the marginal costs will decrease. We will present some sensitivity analyses later to show what impact this may have.

Apart from the development of robots at Statistics Netherlands we have also asked three external companies to submit offers for three airline robots (KLM, Ryanair,

Easyjet). Based on their rates we have also made some statements about the cost-efficiency. For obvious reasons the results of the individual companies cannot be shown here, but the general conclusion is that the hosting costs (which also guarantees that the robot will be updated if a website changes) is actually more expensive than the costs of the manual production process. This means that, under these initial settings, these robots will never lead to costs efficiencies. Also one of the commercial parties offered programming development costs that far exceed the costs of programming the robots at Statistics Netherlands.

As we noted, these calculations are useful for a broad brush view of the efficiency gains (or lack thereof) of internetrobots. The calculations are however very dependent on the assumptions made and the situations included. To illustrate what impact these assumptions have on the results we present a sensitivity analysis in this section (see Table 5). The following variables are altered:

*Internal Robots*

*S1.* Reduce development costs. In this scenario development time of all robots are reduced to a week. In other words, we are assuming that, through learning-by-doing, we can produce new robots in 40 hours. Note that this is still 18 hours quicker than the quickest robot that we have produced so far (easyjet).

*S2.* Halving costs of site changes. Assume that the time required for changing websites halves from 2 hours per month to 1 hour.

*S3.* Combination of S1 and S3. Assume that both S1 and S2 occur simultaneously.

*S4.* Doubling costs of site changes. This is a pessimistic scenario in which we assume that the time needed to reprogram a robot is actually 4 hours per month instead of 2 hours.

*External Robots*

*S5.* The hosting costs of the companies are reduced by 50%. Here we are assuming that it will be possible to renegotiate the hosting costs to 50%.

*Table 5. Sensitivity analysis of break-even points*

| | | **Years** | |
|---|---|---|---|
| | | best case | worst case |
| **Benchmark** | **Scenario** | 3,4 | No BE |
| **Internal robots** | | | |
| Optimistic | S1: Reduce development time | 1,7 | no |
| | S2: Halving costs of site changes | 2,3 | 21,9 |
| | S3: S1+ S2 | 1,1 | 3,3 |
| Pessimistic | S4: Doubling costs of site change | No BE | No BE |
| **External robots** | | | |
| Optimistic | S5: Reduce hosting costs by 50% | 39,6 | No BE |

No BE – No break-even point

Just like the original benchmark calculations the sensitivity analyses show an enormous range of outcomes. The calculations of the internal robots are particularly sensitive to the assumptions about the time to reprogram the robot when a site changes.

A striking result is that if the hosting costs of the external robots are reduced by 50%, the break-even scores are still very poor. Only one company produced a positive break-even score, but this will only come in 40 years.

Some last important comments are in order. The above calculations compare the robots to the existing manual data collection processes. It is important to realize that there is causality between the data collection method and the methodology chosen. For example, the choice of 5 carriers and 5 destinations was motivated by the fact that the CPI knew that the data collection process was manual. In this light it is not surprising that the manual data collection process scores relatively well in cost-efficiency scores because it fits the chosen methodology better.

Note that we have not compared the internetrobots to other optional data collection processes. For example, the CPI may obtain airfare data directly from airline companies. This may also be useful from a methodological point of view because the companies could also provide data on sales volumes. However, it is beyond the scope of our research to also compare the robot to all options available.

Furthermore, the calculations focus on the costs and not on other advantages, particularly the fact that a robot is capable of more extensive data collection which may lead to quality improvements or the solution to methodological issues such as the measurement of the price index for air fares (see figure 1).

Finally, it is also important to realize that for the new applications (unmanned petrol stations and housing market), a manual data collection is by no means a realistic option.

## 5. Conclusions and next steps

A number of conclusions may be drawn from our first experiences of automatically collecting data from web sources. Keep in mind that the main research questions of this project where the *feasibility* and *desirability* of this type of data collection.

Feasibility

1. *Automated data collection is technically feasible but requires special skills*. Internetrobots are applied on the internet frequently. In this report we have shown that this technology can also be applied to the production of official statistics. There are a number of varieties of technology (scripting languages, search engines and robot tools) and Statistics Netherlands may

chose to implement the robots themselves or to outsource this work to specialised external parties. If NSI's produce the robots themselves it is important to be aware that the skills and software required are not part of the regular portfolio of IT-departments.

2. *Use of web-based databases is probably legal.* Based on the database-law and the CBS-law it is probably legal to obtain data through internetrobots. However, since this is not yet a definitive answer we conclude that further legal advice is required. In any case it seems prudent to inform the companies about our activities as a matter of netiquette.

3. *Methodological validity has to be assessed on a case-by-case basis.* Each case which is implemented has its own methodological issues. Our advice is to assess these issues on a case-by-case basis. A checklist, such as the one used by Ossen *et al* (2010) might be of help here.

Desirability

4. *For our case studies, replacing existing manual data collection does not lead to large efficiency gains* The cost-effectiveness calculations for internetrobots exhibit a broad range of outcomes, but are not encouraging. The sensitivity analysis also shows that the assumptions are very influential. Overall one might say that the cost of reprogramming when website changes and the initial costs of the manual data collection are the most important factors in the efficiency score.

5. *Scale and learning by doing will improve cost effectiveness.* It seems incontrovertible that learning by doing and the scale of implementation will lead to large improvements in the programming speed. This is because parts of the scripts can be recycled and debugging experiences reused. Scale effects (i.e. implementing many robots) may also help to renegotiate the contracts of the external parties.

6. *Most potential lies in new applications or more extension data collection.* The internetrobot could be most useful in new areas where the amount of data which is downloaded is beyond the scope of manual data collection. Most benefits seem to lie in the more extensive data collection, which allows for more frequent, quicker and more detailed data. This may lead to quality improvements and improved methodologies. There are a number of new applications fields, particularly the housing market, for which manual data collection is not an option. It is here where the real comparative advantages of the internetrobot lies.

Next steps

This report describes the actions which were carried out in phase one of the project. In phase two of this project, which is already underway, we have selected two external parties to produce and run a number of robots over a period of 3 months (April 1st – July 1st 2010). We will then compare the results and experiences with those of our own (internal) robots. Conclusions on a future approach will be drawn and recommendation for implementation of this type of data collection at CBS will be made. A report on the phase two results is expected at the end of the summer of 2010.

These activities will take place under the auspices of the "Impact of ICT on society" program which was started at Statistics Netherlands at the beginning of 2010.

## References


1. Daas, P. en Beukenhorst, D. 2008. *Databronnen van het CBS: primaire en secundaire bronnen*. DMK notitie.
2. Dialogic, 2008. Go with the dataflow! Analysing internet as a data source (IaD)
3. Roos, M., P. Daas en M. Puts, 2009. Waarnemingsinnovatie. Nieuwe bronnen en mogelijkheden. Report DMH-2008-06-11-PDAS, Statistics Netherlands, Heerlen..
4. Ossen, S., P. Daas en M. Puts, 2010. Automated Data Collection from Web Sources for Official Statistics: First Experiences. Statistics Netherlands.