Automatic data collection on the Internet (web scraping)

Ingolf Boettcher (ingolf.boettcher@statistik.gv.at)[1] *VERSION 18 May 2015*

Keywords: web scraping, Price Statistics, Internet as data source, data collection methods

1. INTRODUCTION

Currently, Statistical Institutes staff members manually collect already a significant amount of data on the internet. The growing importance of online trading requires even more price collection from the internet. Budgetary constraints, however, call for a more efficient deployment of existing human resources to master the additional work load. **Automatic price collection** can, at the same time, support this objective and achieve higher quality price statistics. Nevertheless, legal (permission to crawl on private websites), technological (increased exposure of IT-system to potentially dangerous internet contents), human (need for IT-training), budgetary (implementing and maintenance costs) issues need to be taken into account when deciding on using this new data collection method.

The usage of automatic price collection on the internet as a new data collection method is a two-year project of Statistics Austria's Consumer Price Statistics department. The project is supported by a Eurostat Grant and part of the activities to modernise price statistics [1].

Problem Statement

Every web crawling project poses organizational and methodological challenges to producers of price indices. To put it short: shall a new technology like web scraping evolve or revolutionize the way price collection and index compilation is conducted by National Statistical Offices (NSIs)?

Most NSIs with web crawling projects have opted for approaches that include extensive programming activities, hereby out-sourcing the task of price data collection from Price Index departments. Usually, this approach makes it necessary to re-organize data cleaning, editing and matching procedures. In addition, relatively frequent website changes require re-programming of the web crawler by skilled IT programmers.

The Austrian web crawling project takes advantage of existing commercial visual/click-and-point web crawlers from external companies which allow the development of web crawlers without programming skills. The objective is to keep price collection within the competence of the price index department. Web scrapers should be developed and maintained by price index staff. They are best qualified to apply the various rules concerning the correct collection of price observations and to assess and react on the frequent changes of website. A disadvantage of using click-and-point web crawlers is the dependence on the continuous provision of the external software. Also, at least in the beginning, web crawlers developed by price index staff may not take full advantage of the technology's potential. This is because price index staff usually prefers only an extraction of a limited amount of automatically extracted data, similar to the data collected manually because this allows for an integration of the new method into existing data cleaning, editing and matching processes.

[1] Statistics Austria – Consumer Price Index

2. HOW TO ORGANIZE WEB CRAWLING IN A NATIONAL STATISTICAL INSTITUTE

Web scraping technology offers a wide range of options and may serve different goals: The minimum requirement of a web crawler is to automize the usually manual work of collecting price quotes and article information from websites. The maximum requirement of a web crawler would be to explore price data sources previously unavailable and to provide a census of all price information available on the internet. The decisions made to set up web crawling for price statistics have important methodological and organizational impacts. In general, any price collection procedure with web crawlers will be comprised of at least two steps: data extraction from website and the import of the extracted and validated price data to a data base. Price collection will be followed by cleaning and editing the data and a matching process to price relatives of the previous price collection period.

2.1. Existing NSI web-crawling projects

A web crawler for price statistics needs to structure the data on webpages into rows and columns and extract all relevant information of a product. In order to do so, the crawler needs to be taught how navigate through a given website and how to locate needed data. It has to take into account the individual architecture a website may have and specific features that might require advanced programming such as 'infinite scroll' and JavaScript navigation.

Several Statistical Offices have started projects to use web scraping techniques for their price collection processes. In Europe, Eurostat has supported the initiation of web scraping projects in the NSO's of several EU Member States (Netherlands, Germany, Italy, Luxembourg, Norway, Sweden, Austria, Belgium, Finland and Slovenia).

Of these countries, Germany, Italy and Netherlands and United Kingdom have circulated first results [2]:

Germany and Italy use a methodology that combine web scraping software (i Macros) with java programming to input, select, delete and store data within the price collection process. The Dutch have set up an own web crawling/robot framework using the software R. The British are about to program own web scrapers using the software Python.

The mentioned existing web scraping projects have in common that the development of data collection processes are out-sourced from the price index department to other units qualified to perform necessary programming and data managing tasks. Also, data validation, cleaning, editing and matching procedures are out-sourced as well as the new technology leads to quantitative data sets that cannot be handled any more using existing processes within the price index department.

[2] For an overview on the German, Italien and Dutch project see: Destatis. Multipurpose Price Statistics Objective D: The use and analysis of prices collected on internet Final Report March 2014

2.2. The Austrian Approach – Using visual/point-and-click Web Crawler

In contrast to other web scraping project, Statistics Austria has opted to deploy point-and-click "visual" web-crawlers only to avoid any programming activities. Several considerations have led to this decision:

Visual web scrapers are available at very low or no costs at all

Recently, the availability of visual/point-and-click web crawlers has increased substantially. Before, web scrapers had to be developed writing custom code by skilled IT personnel. Examples of visual web scrapers are outwithhub and import.io. Statistics Austria uses import.io for its pilot web scraping project. However, the dependency on a single commercial software provider should be avoided by NSIs. Therefore, any web scraping project using external software is well advised to prepare for cases that require the replacement of the used software, such as service denials, business closures or other reasons.

Low implementation costs

Visual web scrapers substantially reduce the necessary programming skills. IT costs are limited to setting up the web scraping environment in accordance to IT security requirements. Click-and-point web scrapers that only imitate existing price collection practice can be developed in a few minutes (e.g. the collection of prices for all product on a specific website's URL).

Low intra-institutional and organizational burdens

Visual/click-and-point web scrapers can be developed and maintained by the price collection team themselves. Other departments (IT, data collection, methods) do not need to take over new tasks and employ new personnel, respectively. In the beginning, only the data collection processes are affected. Existing data validation, cleaning and editing processes don't need to be changed (as long as the automatically extracted data has the same features as the manually collected data). [3]
In the long run, to increase price index quality, the full use of web scraping potential will lead to an exponential growth of price data information that will require adapted data cleaning and editing processes (e.g. if the collection of flight prices is extended to daily data collections instead of once a week).

[3] The advantage of avoiding immediate changes to reliable data quality assurance processes should not be underestimated. Web crawlers can create a large amount of price observations that may be used to create price indices of superior quality. However, price collection is not the only task within the process of price index compilation. Price observations need to be matched to previous periods and to product categories and might need to be weighted. For example, the collection of price data on all existing flights from Austrian territory will be easily available using web crawlers. However, in order to use all the flight price information reliable weights for every flight destination need to be compiled using recent airport statistics, hereby substantially increasing the work load for another step of the price index compilation.

Higher flexibility

Existing web crawling projects have reported that websites change frequently which requires the re-programming of the respective web crawlers. It may be expected that changes to websites will be more easily spotted and immediately taken care of by directly responsible price index team members than by IT personnel. External website changes might also require external programming know-how unavailable to internal IT personnel. There are a variety of innovative website features coming up that need advanced web crawling solutions, such as 'infinitive scroll' which replaces pagination on some retailer websites. The developers of external visual web scrapers identify these innovations and develop solutions.

The emergence of the data scientist

In the long run, national statistical offices will need to produce more and higher quality statistics with fewer resources. It seems unreasonable to employ skilled and expensive IT personnel frequently for relatively simple tasks such as adapting a web scraper to website changes by re-writing custom code.

At the same time, the education and qualification of personnel within the price index teams will improve due to demographic reason. Manual price collection is not an attractive task to do for well trained staff. Having the responsibility to develop, run and maintain an automatic data collection tool, however, is a more qualified work task. Instead of out-sourcing complex IT tasks, price index team members can develop their skill and develop to data scientist, hereby bridging the existing gap between Statisticians and IT personnel in National Statistical Institutes.

3. STATISTICS AUSTRIA – WEB CRAWLING PROJECT OUTLINE

The project is structured as follows:

- *Selection of software*

The automatic price collection software has been selected according to several criteria.

- The software must provide a high level of *usability*.
- It has to be software that can be easily understood by non-IT price statistics staff members.
- The software should provide a surface that enables users with basic IT knowledge to change the price collection procedure (e.g. in case of website changes).
- The software must provide a well maintained documentation and should be adoptable the internal IT system.
- Before any implementation, IT-specialists assure that the software is safe to operate and that it comes along with appropriate licensing, testability and supportability.
- A risk analysis assesses the potential legal and data security problems.

-Legal Analysis

The legal department assesses the national legal framework concerning the jurisdiction on the extraction of online product information for statistical purposes. As a result, the legal requirements are taken note of and a stringent 'rules of conduct' for the automatic price collection have to be written and published which transparently describes the methods used to perform web scraping.

- *Implementation and maintenance of software and supporting IT infrastructure*
The IT department acquires and installs the selected software. Maintenance procedures to update and test the software regularly and to provide support needs to be set up and documented. Automatic web scraping has been identified as a potential leak for Statistics Austria's IT-System. In order to avoid viruses, hackers etc. to infiltrate the internal IT-system the web scraper operates within a standalone system on a separate server. Employees access the software and the scraped data by using a remote server from their PC. Furthermore, IT develops and maintains an infrastructure (SQL Database) to store the extracted data.

- *Selection of Product Groups and Online Retailers*

In the beginning, Product Groups and Online Retailers are selected according to currently valid manual price collection procedure. This approach facilitates the comparison of the results from automation. In a later step, product groups and retailers not yet in the price index sample will be targeted.

-*Development of automatic price collection processes using the selected software*
Price statistics staff use the web scraping software and create automation scripts to continuously download price data from eligible online retailers. This step includes checking the compatibility of the specific extraction methods applied on the selected data-sources (online retailers). *Quantitative* as well as *imitating approaches* are considered. The Quantitative approach aims at continuously harvesting all the available price data from selected websites. The imitative approach collects automatically the data according to criteria, which are currently already applied in the manual price collection. Internet data sources are connected directly to output files (e.g. live databases and reports), the extracted data is analysed and cleaned for price index compilation. In the end, an automatic price collection system will produce data that can be directly used for the production of elementary aggregate price indices. Quality control and price collection supervision as well as changes to the automation scripts are done by price statistics department staff. IT infrastructure and software maintenance (updates) are supplied by IT.

-*Development of quality assurance methods*

Part of the quality assurance is the comparison of automatically collected price data with manually collected prices. Later, predefined research routines and consistency checks will be deployed. An optimal method would be the deployment of second web crawler software whose results could be automatically compared with the results of the first web crawler.

-*Usage of automatic price collection for various price statistics*

In order to maximise the output of the investment into automatic price collection, the actions of the project will aim at the inclusion of as many price statistics as possible. Thus, all price statistics projects will cooperate on the development, in particular HICP and PPP, but also other price statistics such as the Price Index on Producer Durables.

4. DESCRIPTION OF VISUAL/POINT-AND- WEB SCRAPING SOFTWARE[4]

A visual web scraper relies on the user training a piece of crawler technology, which then follows patterns in semi-structured data sources. The dominant method for teaching a visual crawler is by highlighting data in a browser and training columns. [5]

[4] http://support.import.io/knowledgebase/articles/251954-what-are-extractors-crawlers-connectors

[5] See also the respective wikipedia article on web scrapers: http://en.wikipedia.org/wiki/Web_crawler

Visual web scrapers are able to operate because most websites are built and structured in a similar way. Click-and-point software solutions make it possible to create an algorithm to handle the individual elements of a website. Writing custom code by a dedicated programmer becomes unnecessary.
In order to deal with different webpage architectures and the individual data needs different scraping tools are offered. The software used by Statistics Austria calls these tool *Extractors*, *Crawlers* and *Connectors*. Together these three tools allow for all kinds of data scraping activities.

Extractors

Extractors are the most straightforward method of getting data. In essence, an extractor turns one web page into a table of data. In order to do this, the data is mapped on the page by highlighting it and defining it. If there is a pattern to the data, the extraction algorithms will recognize it and pull all your data into a table.

After the creation of an extractor, price data collectors can refresh the data whenever needed. Once a website has been mapped, data from a similar page can be extracted without mapping anything.

Example:
Extractors are used in the Austrian project to extract data for electronics and clothing. The extractor has to be built only once for a certain webpage. After that the responsible price collection team member identifies all the URLs for the specific product groups collected from the online retailer (e.g. Notebooks, Mobile phones, Smartphones, PCs, etc.), then copy pastes the URL into the import.io application and runs the program. A pagination function makes sure that all the available data is extracted.

Connector

A connector is an extractor that is connected to a search box. Connectors allow to record your searches on a website and to extract the data for that search. Connectors, unlike Extractors and Crawlers, interact with sites, such as search boxes, clicks, options and inputs like logins.

Example:
Connectors are currently used in the Austrian project to extract data for flights. The extractor has to be built only once for a certain webpage. After that the responsible price collection team member adapts the connector for all measured flight connections. He/she also enters the required flight dates.

Crawlers

Crawlers allow to navigate through all the pages of a website and to extract all of the data on every page that matches the pattern you mapped.
For example if the product details for every item of a retailer are needed, an extractor for one product page would have to be mapped. However, instead of restricting the

extraction to certain URLs the crawler searches the site for all the other product pages similar to the mapped one and extracts the chosen data.

Data upload

Collected data is first uploaded to import.io's cloud servers. The application allows to download the scraped data as a file. CSV (comma separated values), TSV (tab separated values), XML (Extensible Markup Language), JSON (JavaScript Object Notation) file formats are supported. CSV files can be opened using any spreadsheet software like Microsoft Excel, Open Office or with Google Docs.

5. RESULTS

Currently, the pilot project performs all project tasks using the web crawling software *import.to*. The main advantage of the tested software is that no advanced programming skills are needed to perform changes to the web crawling programs in case of website changes.
The success of automatic data collection depends on the ability of the deployed web crawler to simultaneously improve the data quality while reducing the overall data collection costs. Table 1 provides first details on the ability of automatic price collection to achieve these goals :

Table 1. Comparison – Manual vs. Automatic Price Collection method - Flights

Method	Product Group	# of Prices	Work load	Comment
Manual	Flights	Ca. 200	16h per month	Ca. 5 min per price
Web crawler	Flights	Ca. 4000	4h+X	X= irregular maintenance work

The monthly working hours spent to collect the prices needed to compile the index for prices on passenger flights can be substantially reduced from 16 to 2 hours. In fact, the actual manual price collection has been completely replaced and the quality of the price index will be higher due to an increased number of measured price quotes. The two hours needed for the automatic price collection method cover various tasks, such as data importing, data cleaning and data checking. In the course of the project, the work load factor X, the irregular maintenance work needed to run the web crawler, has to be assessed and quantified. Maintenance is required when website architecture is changed. There is evidence that the resources needed to perform the irregular maintenance work depends on the individual website and heavily affects the total work load. Thus, a critical cost effectiveness analysis is needed when applying automatic price collection methods.

6. LEGAL ASPECTS

Web scraping techniques used within an automatic data collection project should always be checked against any legal restrictions imposed by the legislator. In Austria, there have

not been yet any legal proceedings concerning the admissibility of web scraping. However, in other European countries, such as Germany, there have been already court decisions on the rights of online database owners to prevent web crawlers from systematic copying of their content.[6]

Statistics Austria's legal department thoroughly researched the available legislations and court decisions and interpretations on the topic. It found that the use of web crawlers for official statistics is legal under certain conditions. An entrepreneur, who places an offer on the internet accessible to the public, must tolerate it that his data is found and downloaded by conventional search services in an automated process. The conditions any application of a web crawler should adhere to are as follows:

Technical hurdles of websites may not be circumvented.
There are technical solutions available for website builders to block or delay web crawlers (Passwords, robot blockers and delays in a websites robot.txt – file, etc.). Such techniques should be respected by web crawlers designed by National Statistical Offices for automatic data collection on the internet.

The database may not be replicated as a whole elsewhere through web crawling
The crawling of a database shall not cause any damage to the owner. Thus, a simple replication of a websites full content is not allowed as this would create a direct competitor.

Web crawling may not negatively affect a web sites performance
Web crawling technology has the potential to reduce the performance of a websites. The number of executions caused by the web crawler on a website should be as low as possible. Especially the frequency of executions should be set at an absorbable rate for any professionally design website (e.g. max. 10 executions per second).

7. CONCLUSIONS

The web crawling technology provides for an opportunity to improve statistical data quality and to reduce the overall workload for data collection. Using automatic price collection methods enables statisticians to react better to the increasing amount of data sources on the internet. Any implementation of the method requires thorough planning in various fields. Legal and data security aspects need to be dealt with in the beginning. Necessary IT resources and IT training required to maintain the automatic data collection system have to be estimated in the course of a pilot project and should not be under-estimated.

[6] Brunner, K. & Burg F. (2013) DESTATIS Statistisches Bundesamt Multipurpose Price Statistics. Objective D: The use and analysis of prices collected on internet. Interim Report, P.9

REFERENCES

[1] R. Barcellan, Multipurpose Price Statistics, Ottawa Group (2013), 9. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/8bdac0e73d96c891ca257bb00002fdb4/$FILE/OG%202013%20Barcellan%20-%20Multipurpose%20Price%20Statistics.pdf

[2] Destatis. Multipurpose Price Statistics Objective D: The use and analysis of prices collected on internet Final Report March 2014

[3] Web scraping techniques to collect data on consumer electronics and airfares for Italien HICP compilation. ISTAT. 2014. http://www.q2014.at/fileadmin/user_upload/Iad_in_ICT_survey_PAPER.pdf

[4] On the use of internet robots for official statistics. CBS. 2014. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf