From price collection to price data analytics

How new large data sources require price statisticians to re-think their index compilation procedures.
Experiences from web-scraped and scanner data

Josef Auer and Ingolf Boettcher[1]

***Statistics Austria** – Consumer Price Index*

Abstract:

New data sources such as web-scraped data and business transaction data (e.g. online and scanner data from retailers for price statistics) have the potential to improve official price statistics, both in terms of quality (more data) and efficiency (low data collection costs, lower response burden). However, the integration of new data sources in price statistics is often not straightforward. Statisticians have to review and eventually replace traditional price index compilation procedures to comply with existing quality standards. The challenges to deal with are manifold: How to combine traditional and often primary data sources with new large, secondary and often raw data sources? How to assess the representativeness of data sets from the internet for official data production? How to validate integrity and completeness of scanner and web-scraped data? How to integrate new and diverging data set structures into well-established statistical production processes?

[1] STATISTICS AUSTRIA, Directorate Macro-economic, Statistics Prices and Purchasing Power Parities
Email: josef.auer@statistik.gv.at and ingolf.boettcher@statistik.gv.at

The opinions expressed in this paper are the author's own and do not necessarily reflect the view of Statistics Austria.

Parts of this paper are adaptions from the author's conference paper 'Quality Control of Web-Scraped and Transaction Data' presented at the Q2016 in May 2016 and from Statistics Austria's EU- Grant Reports on its Scannerdata and Web-scraping projects (both 2016).

INTRODUCTION

Statistics Austria deploys several price collection methods to compile consumer price indices. Price collection is organized centrally via email, fax, internet and telephone, and regionally via price collection in actual outlets. Two major developments make it necessary to modernize existing price collection methods:
Firstly, more and more consumer products are sold using flexible and/or individual pricing schemes (promotion prices, membership prices). Conventionally measured list prices are becoming less reliable. The actual prices paid are captured only in the retailer's transaction data sets (e.g. scanner data from supermarket chains).
Secondly, the growing importance of online commerce: As regards e-commerce, conventionally measured list prices become less reliable as online pricing schemes are highly flexible. Website prices fluctuate significantly depending on increasingly complex price setting algorithms. Prices may depend on time / place (IP-Address) / identity (member vs. non-member), quantity, demand history, etc. (This is the case in particular for airfares and hotels, less so for food, clothing and electronics).

Both developments require official statistics to adapt data collection sources and methods in order to maintain high statistical quality standards. Despite all difficulties, they are an opportunity to improve the quality of official price statistics. This is because, in comparison to conventional data sources, transaction and online data potentially allow total coverage of the target universe (e.g. price paid for consumer products) along several dimensions: time (every day), products (all items) and markets (all stores). Also, data collection efficiency increases as transaction and online data allow a shift from manually collected prices to automated data collection processes. In addition the response burden for businesses is reduced.

Statistics Austria reacts on the growing importance of transaction and online data by conducting several pilot projects on the use of scanner data and web-scraping, supported by Eurostat.

The Austrian Scanner Data and Web-Scraping Projects

The Austrian scanner data project currently focuses on retail segments offering standardized consumer goods which are relatively easy to categorize and to process for CPI compilation. Food and beverages are the most important product groups for the scanner data project. Other possible retail segments are drugstores, medical and pharmaceutical products, non-durable household goods, small tools, accessories for cars and miscellaneous accessories.

Mostly, these segments are currently covered by regional price collection in 20 Austrian CPI regions. Durable goods (e.g. cars) and major appliances (e.g. washing machines) are for the time being not targeted by the scanner data project. These product groups often require quality adjustments according to international standards and are hard to automate and continue to be implemented by trained CPI staff at the central office.

Altogether about 20% of the Austrian CPI basket of goods may be covered using scanner data.

The Austrian web-scraping project focuses on products and services for which price data is currently manually collected on the internet. At the moment, the most important target segments for web-scraping are transportation (e.g. flight tickets, train tickets, holiday package tours), technical equipment, clothing and hotels.

With the exception of ‘clothing ’, these segments are mainly covered by central price collection. Momentarily, altogether about 10% of the Austrian CPI basket of goods may be covered using web-scraping.

The challenge to shift from price collection to price data analytics

The introduction of new data sources such as scanner data and web scraping has a profound effect on price index compilation procedures. The main reason for this is that new large price data sources are secondary data sources while price index compilation procedures usually are based on primary data sources. Statistics based on primary data sources are characterized by a high workload for supervising and conducting surveys. In general, price index staff at the National Statistical Institutes is specialized on obtaining structured price information rather than processing unstructured price data. Providers of price data may be retail outlets, company head offices, authorities, households, etc. Conventional data sources include outlet, catalogues, websites, etc., all of them having in common that price data input is structured according to pre-defined survey forms.

Shifting to price data analytics by using large price data sources adds several steps to existing price collection processes. A sheer unlimited source of available IT solutions and applications for data analytics supports this task. This includes a range of technologies often including, but not limited to:

-Big data platforms (i.e. Hadoop and Spark), analytics engines, and programming languages (like Python, SAS and R).
- Visualization, and reporting applications.
- Data warehouses and databases (for example, Cassandra).
- Security frameworks.
- Web crawling tools (web tools, part of statistical software packages)
- Servers, storage infrastructure.

A significant problem is that the integration of price data analytics into a price index compilation process by using elaborated analytics technologies is a comprehensive task that needs a broad and resourceful environment. The resources necessary to plan, develop, test, deploy and manage a reliable price data analytics solution should not be underestimated. The range of new skills to be developed can be wide; depending on how much new data analytics technologies are being used. Without initial additional resources price index departments might find it hard to implement new large data sources, such as web-scraped price data and scanner data, into their price index compilation processes. In this regard, Statistics Austria has profited from EU-funded Eurostat grants that support the implementation of new technical solutions for data sources and data validation.

This paper will demonstrate the challenges when using scanner and web-scraped data for CPI compilation. Price index compilation procedures have always been based heavily on traditional price survey methods. However, in light of budgetary constraints on the part of the NSIs and increasingly flexible pricing schemes and transnational trade on the part of retailers, traditional price surveys become more and more inappropriate for application in several product and service segments. To compensate for these developments, price statisticians have to work out new methods and procedures that allow the efficient compilation of high quality price indices, namely by filtering, manipulating and examining price information from vast and large, raw data sources. Specific examples in the annexes describe how the paradigm shift in price statistics steaming from the use of large new data sources can be handled.

Chapter 1 and 2 of this paper describes how the implementation of large new data sources require a new assessment of data quality and the introduction of new data quality control processes . Chapter 3 describes some principles to consider when introducing web scraping for CPI compilation purposes. Annex 1-4 provide some detailed insights to specific methodological problems and data processing solutions when dealing with scanner data and web scraped data, respectively (Assigning scanner data sets to CPI basket items, introduction of new strata, milestone to introduce new large data sources, handling of irregular input data).

1 Assuring data quality of large new data sources

Compilers of official price statistics deploy reliable quality assurance processes that have been improved for years. However, statisticians who want to integrate new sources such as web scraped and scanner data into official statistics will be faced with a variety of challenges to cope with. Anybody dealing with one of the new large data sources is not just faced with replacing one source of data with another. Instead, new large data sources require a comprehensive re-thinking of the whole statistical production process.

Approach to data quality assurance of large new data sources

No matter what social or economic phenomena is looked at, official statisticians usually must define a universe /an entire statistical population of a phenomena, then classify its characteristics, measure statistical data and turn it into meaningful statistical information.

In the age before big data, it is mostly not possible to measure every single element of the statistical population. Instead, statisticians often work with samples that are representative for the whole statistical population. Now, with large new data sources being available at almost no costs it appears that official statistics has potentially access to the universe of a statistical phenomenon such as consumer goods prices. It might seem that one just has to turn all available prices into a price index after applying some input quality processes. Statistical models might become simpler as it can be argued that the available data represents the entire universe, e.g. the totality of prices. At least it might look like that and in fact, it is compelling to just accept the new large data sources as correctly mirroring the universe that allow us to skip the whole statistical sampling step altogether. But no matter what a statistician is doing with large new data sources: the control of the price data's relevance, completeness and accuracy has to be part of any statistical production process.

New big data sources come with the promise of mirroring reality by featuring not a sample but the totality of the statistical population that is subject to statistical work. However, the question on how to process large new data sources is not just about cleaning up the data. It must be kept in mind that most of the time, and surely for price statistical data, no data source represents the totality of the statistical population. There is always some form of bias, over- or underrepresentation. Therefore, access to new large data sources does not replace basic methodological work and checks concerning coverage bias, measurement error and self-selection bias. New kinds of processes and approaches to data quality assessment, as exemplified below for web-scraped data, have to check upon the completeness, accuracy and consistency of every data set:

Data quality assessment of web-scraped price data (exemplary)

Coverage testing

Coverage is about testing if the automatically extracted website data has the right amount of records. Every extraction should go along with a number that sets a benchmark for extraction. In case price collection for CPI the number of records should be equal or close to the number of records in the previous month. Any serious deviation from that benchmark should be monitored and checked upon.

Completeness testing

Just because roughly the right number of records were extracted does not actually mean that the right number of results (cells with the right data in them) have been returned. Completeness is about measuring which columns should always have something in them (e.g. every item should have a price), which should sometimes have something in them (e.g. some items will have a promotion price).

How important completeness is depends on what is done with the data. If the variables are used to plot equivalent room categories for hotel prices after data exporting, all the product information would be essential. If the products segment is very homogenous only some variables might be essential (e.g. product ID for matching purpose). In any case, if an web extractor returns a lot of missing fields, responsible quality control triggers processes that include going back and looking at the data source to make sure the information really is on the page and –if it is –retrain the web data extractor to deliver more complete results. Sometimes if the page structure changes slightly from result to result, there is a need to do a little extra training to expand the underlying scrape algorithms.

Data type testing

Each property (column field) of your data set has a certain type – which is defined when creating an extractor output table (by naming the variables and marking the specific data field locations). In addition to being simple things like number, text or currency; types can also define validation rules, such as expected string patterns. For example, zip codes contain four or five digits (optionally separated by a hyphen) or a price is valid if it consists of a number and possibly a currency sign.

Manual testing

Periodical manual testing of automatically extracted web data should stay a standard method to check and document data quality. In general, it should be enough to manually check about 10 records (rows of data) against the source website, hereby verifying whether the extracted data in the exported spreadsheet is the same as the data on the page. This remains a reliable way to spot any obvious problems when automatically extracting price data.

2 New quality control approaches for transaction and web-scraped data

New data sources such as transaction and web-scraped data are subject to existing quality control standards. However, there is a lack of applicable quality control procedures and guidelines regarding data from large and vast secondary data sources (transaction data, web-scraped data, social media data, internet of things data, etc.). Compilers of official statistics will find it hard to apply existing data quality frameworks to large data sources. In fact, official statistics quality frameworks have for a very long time focused on primary statistical data sources (e.g. ESS quality report 2014). In these quality frameworks the output requirements of official statistics are thoroughly described. In the last years, updated quality frameworks focus more and more on the integration of secondary non-statistical data. In particular the integration of administrative secondary non-statistical data for official statistics has been in the spotlight of quality frameworks (see UNECE 2011). Recently, works have started to provide guidelines and frameworks for the quality of input data for official statistics users to guide statistician when compiling official statistics from Big Data sources (Eurostat 2017; Struijs P. and Daas P. 2014).

Table 1 and 2 below depict several novel challenges encountered by the Austrian scanner data and web-scraping pilot project. The tables contain the most relevant quality criteria for input data and only depict problems and measurement methods that are unique to transaction and web-scraped data – while quality problems known when dealing with conventional primary or secondary data sources are not treated.

Table 1 – Novel quality problems and measurement methods with transaction data

Input data quality criteria	**Transaction data /scanner data**	
	Novel quality problem (for consumer price statistics)	**Measurement Method**
Relevance	Data may contain transactions that are out of scope. -e.g. expenditures for business purposes (out of scope for consumer price indices)	Information by data providers; otherwise unresolved
Accuracy	Volume and variety of data sets are too large to identify and clean erroneous/ untrustworthy/ inconsistent data sets with conventional methods.	Extent in % of erroneous / inconsistent data is monitored and excluded
Timeliness/Punctuality **Accessibility**	- (no new kind of quality problem)	-divergence from formal data delivery agreement between data provider and NSI
Completeness	Volume and variety of data sets are too large to identify missing values with conventional methods. (Scanner data: natural attrition of Unique identifiers is extremely high)	Number and level of target values are measured against historical values from previous deliveries
Clarity / interpretability	(no new kind of quality problem)	Information by data providers about: -format and definition of variables -data transformation checks performed before delivery (e.g. aggregation)

Challenges of the use of scanner data

Table 1 shows that the use of transaction / scanner data poses several quality challenges that need to be addressed. Replacing traditional price collection with scanner data leads to a high dependency of Statistical Offices on the providing retailers. Therefore, the quality provisions of the delivered transaction data should preferably be laid down in a formal agreement. Scanner data is a secondary data source and may include data types, classifications, characteristics and elements that are hard to integrate with the existing CPI production system. Processing scanner data can be difficult as each retailer usually deploys different database structures, data types and product classifications. Extensive data cleaning and index compilation procedures need to be developed for each scanner data provider. In particular, there are two main tasks to achieve when processing scanner data after receiving it from retailers and before CPI compilation: matching individual articles between time periods and assigning/mapping GTINs to a CPI elementary aggregate (EA) and COICOP (sub-)class (see also annex 1). Also, size and structure of the data files might require investments in IT infrastructure. Finally, scanner data raise methodological issues that need to be addressed to ensure that existing rules establishing comparable CPIs in the EU are not violated.

Table 2 - Novel quality problems and measurement methods with web-scraped data

Input data quality criteria	Web-scraped data	
	Novel quality problem (for consumer price statistics)	**Measurement Method**
Relevance	Representatives of online data (are products offered really sold and by whom?)	Information by data providers; otherwise unresolved
Accuracy	Website content may be IP-specific (a user who frequently checks a website or a web-scraper might lead to different price displays than first-time users)	Comparison of automatically and manually collected data
Timeliness/Punctuality	the amount of data makes it difficult to judge data quality within a reasonable amount of time	-quantitative instead of qualitative processing of data
Accessibility	Websites might identify web-scrapers and block them	unresolved
Completeness	Websites change frequently Relevant variables and URLs might not be identified and scraped	Number and level of target values are measured against historical values from previous data collection activities
clarity / interpretability	- (no new kind of quality problem)	

Challenges of the use of web-scraped data

The main challenge of the Austrian web-scraping project is the issue of frequently changing websites structures. This requires the re-programming and adaption of the respective web-scrapers. In order to keep the necessary IT resources within reasonable limits when adapting web-scrapers, click-and-point software tools are deployed to develop and maintain the web-scrapers.

Development of automatic price collection quality assurance processes
Price statistics staff uses the web-scraping software to 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. The extracted data is analysed and cleaned for price index compilation.

One part of the quality assurance is the comparison of automatically collected price data with manually collected prices. Predefined research routines and consistency checks are deployed. It would be beneficial to deploy a second web-scraper software whose results could be automatically compared with the results of the first web-scraper. The irregular maintenance work needed to run the web-scraper is significant and should not be underestimated. Maintenance is required to assure quality when a website changes its structure.

Last but not least, issues related to data security must be taken into account when implementing web scrapers within the sensitive IT environment of a National Statistical Institute. Web Scraping is a natural threat to any professional IT infrastructure that puts data security as a priority. Statistics Austria IT-security experts have identified several risks, in particular the potential download and execution of virus-software. Therefore, the decision has been taken to execute web scraping within a stand-alone system. Hereby, viruses and mal-software that have been downloaded through web-scraping cannot infiltrate the internal IT environment of Statistics Austria.

3 USE OF WEB-SCRAPERS FOR OFFICIAL STATISTICS

The decisions made to set up web-scraping for price statistics have important methodological and organizational impacts. In general, any price collection procedure with web-scrapers comprises 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.

A web-scraper 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 scraper needs to be taught how to 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, Netherlands and United Kingdom have circulated first results.[2] Germany and Italy use a methodology that combine web-scraping software (iMacros) with java programming to input, select, delete and store data within the price collection process. The Dutch have set up an own web-scraping/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 source code programming and data managing tasks. Also, data validation, cleaning, editing and matching procedures are out-sourced 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

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-scrapers from systematic copying of their content.[3]

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-scrapers 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-scraper 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-scrapers (Passwords, robot blockers and delays in a websites robot.txt – file, etc.). Such techniques should be respected by web-scrapers designed by Statistics Austria for automatic data collection on the internet.

The database may not be replicated as a whole elsewhere through web-scraping

The scraping 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-scraping may not negatively affect a web sites performance

Web-scraping technology has the potential to reduce the performance of a website. The number of executions caused by the web-scraper on a website should be as low as possible. Therefore, the frequency of executions should be set at an absorbable rate for any professionally design website (e.g. max. 10 executions per second).

CONCLUSION

Regarding the implementation of large new data sources, new kinds of skills ("data science") are required from statisticians to build up processes that comply with existing quality standards on data output. Advanced knowledge of data manipulation and programming is necessary to succeed in this task. Statisticians will have to invest into training of staff within their unit and improve and integrate cooperation with colleagues from other departments who are able to handle large data sources (e.g. IT, data collection departments). All in all, the right amount of statistical creativity is necessary to transform new and ever-changing (big) data sources into high quality official statistics. Large secondary data sources require individual data cleaning and editing processes. Measurement of big data input data quality will make more flexible measurement methods and quality benchmarks

[3] 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

necessary. To facilitate these challenges, the statistical community should continue the work on guidance and quality frameworks for integrating large new secondary data sources into official statistics. This is especially important, as NSIs usually are facing financial constraints. There might be a danger of only using the advantages of new secondary data sources (low collection costs, high coverage) and to neglect the disadvantages . Large new secondary data sources have the potential to improve official statistics but also to cause faulty and biased outcomes as the degree of necessary data manipulation quality related decisions by statisticians increase.

ANNEX 1 - ASSIGNING ARTICLES TO CPI BASKET ITEMS

Considerable resources need to be applied for classifying, coding and mapping scanner data to allow their integration into the CPI compilation processes. There is no ISO standard which requires retailers to structure their point of sales data and internal product classification system in a harmonized way. Instead, most retailers have built up their own systems and classifications which often support the organizational structure and strategy of the retailer. Classification systems might be organized according to marketing and logistics principles, respectively, rather than according to classification principles of product purpose (e.g. in test-scanner data available to Statistics Austria ice cream bars were assigned to different internal product groups of a single retailer called "cash point zone products", "frozen food" and "ice cream"). Depending on the retailer, the structure and characteristics of the scanner data of retailers might be still based on 1980s database conventions with very few characters and digits per variable due to restricted storage place. Thus, a CPI compilation system with scanner data has to cope with diverging data structures of the retailers and scanner data providers.

There are two main tasks to achieve when processing scanner data after receiving it from retailers and before CPI compilation: matching individual articles between time periods and assigning/mapping individual articles (GTINs) to a CPI elementary aggregate (EA) and COICOP (sub-)class.
Concerning the *matching process,* GTINS and internal product codes are the obvious first choice for the correct identification of individual articles in different periods of time. However, one of the major problems concerning scanner data is that GTINs and product codes of articles change over time, e.g. when a product re-launch occurs, the place of production changes, internal changes of the manufacturer take place, etc. Attrition rates of GTINS/Product codes are up to 45% in a one year period. Thus, the Austrian scanner data project works on the automation of processes that support the matching process by taking into account product characteristics when identifying identical or at least homogenous articles over time. Table 3 lists the different options currently used by NSIs for the *assigning process* which maps article scanner data sets to a CPI elementary aggregate and COICOP (sub)-class, respectively. (Please note that the list is neither complete nor absolutely precise as many NSIs use a mixture of methods when working with scanner data.)

Table 3 -Assigning scanner data sets to CPI elementary aggregates / COICOP (sub)-class

Article Mapping Method	Description – Advantages – Disadvantages
Using GTIN Dictionary from Market Research Company	*Description* Merging the scanner data information from retailers with the detailed product characteristics from the GTIN Dictionary by GTIN code. Advantage -low work-load to assign articles to CPI Elementary Aggregates *Disadvantage* -costs to obtain GTIN Dictionary -not all articles use universal GTIN Codes (e.g. private label products that are exclusively sold by retailers) *NSIs applying the method* INSEE France
Harmonizing the classification with retailer	*Description* The retailer develops / changes / streamlines in cooperation with the NSI a new or existing classification in order to comply with the product groups of the NSI and with COICOP. *Advantage* -low work-load to assign articles to CPI Elementary Aggregates *Disadvantage* -willingness of retailer usually low to harmonize own classification *NSIs applying the method* CBS Netherlands
Reducing the Sample Size and manually assigning articles to EAs	*Description* Drawing a representative sample of articles with significant turnover and manually assigning articles to Elementary Aggregates *Advantage* -good overview and understanding of the used data - *Disadvantage* - high manual work load - not making full usage of Scanner data potential *NSIs applying the method* BFS Switzerland; Statistics Denmark; Statistics Sweden
Automatically assigning products using statistical software data management	*Description* Data management that assigns articles to EAs according to existing information (internal group of products, item description etc.). Pre-requirement: Development of key-word list and data management programming. *Advantage* -no need to draw sample from the scanner data census -no need to obtain additional article information from market research companies *Disadvantage* -long implementation time (building up key word lists + re-programming) -high work load needed for quality control and key word list maintenance *NSIs applying the method* Statistics Portugal and Statistics Austria (both test projects)
Acquisition of Scanner Data from market research companies	*Description* Scanner data from market research companies include detailed article classifications. *Advantage* -low work load *Disadvantage* -high costs -Discounters (Aldi, Lidl) not included -no real supervision of data management procedure possible *NSIs applying the method* -currently none

Scanner data contains information on thousands of different products. All articles must be linked to the COICOP classification in the course of CPI compilation. At first, manufacturers own classification is reviewed and checked upon similarities to COICOP. In the best case, one or several retailer classification classes generate a COICIOP and elementary aggregate definition, respectively. However, many retailer classifications classes do not cover perfectly internal item definitions. An initial allocation exercise must therefore be performed. Some retailer classes might be disregarded because of negligible expenditures. The GTIN of other classes need to be allocated manually to the respective CPI elementary aggregates. This represents a high cost, which, however, is a one-time exercise. In consecutive month, only newly arriving GTIN that belong to these retailer classes have to be attributed manually.

Table 4 below displays an example of retailer specific article groups that can be perfectly attributed to CPI Elementary aggregates. The retailer's article group Orange juice and Apple juice can be allocated to the CPI Elementary Aggregate codes 346 and 347. As long as there are no changes of the retailer classification names and codes in the master data set, this allocation can be implemented automatically without any further manual allocation in the future. However, when detecting classification changes manual checks should be done to verify if the automatic allocation is still correct.

Table 4 - Allocation of retailer classes to CPI-Elementary aggregates - automatic

Article segment	**Article class**	**Article group**	**CPI-EA-code**	**CPI-EA-Description (short)**
…	**…**	**…**	…	…
Beverage	Juice	Tomato juice	-	-
Beverage	Juice	Peach juice	-	-
Beverage	Juice	**Orange juice**	346	**Orange juice**
Beverage	Juice	**Apple juice**	347	**Apple juice**
…	…	…	…	…

Tables 5 and 6 below describe the working steps necessary when an automatic allocation of a retailer's article sub-class is not possible. The products of the two diaper - classes of goods "jumbo pack" and "stock package" cannot be clearly assigned to the CPI elementary aggregate 367 " Diapers " as it relates to diapers for babies weighing 3-6kg (regarded as a homogeneous product group). The assignment to the CPI elementary aggregate 367 "diapers for 3-6 kg baby " is carried out manually. All product descriptions of GTIN belonging to that class need to be reviewed and allocated according to the available information.

Table 5 - Allocation of retailer classes to CPI-Elementary aggregates - manual

Article segment	Article class	Article group	CPI-EA-code	CPI-EA-Description (short)
…	**…**	**…**	…	…
Diapers	**Diapers (disposable)**	**Jumbo-pack**	367	Diapers for 3-6 kg-Babies
Diapers	**Diapers (disposable)**	**Stock package**	367	Diapers for 3-6 kg-Babies
…	**…**	**…**	…	…

The retailer's article master data set contains a total of about 130 diaper products which have to be manually assigned in a one-time exercise. For some products the product information is not enough to perform the assignment exercise and further product research has to be conducted, e.g. on the website of distributors or manufacturers. In our example in table 6, it is necessary to research more detailed article descriptions for the product "babylove MegaPack Maxi 126piece". The CPI-elementary aggregate requires the collection of diapers for babies weighing about 3-6kg. The manual research on the internet about the product needs to assess whether the "babylove MegaPack" falls into that category. Such manual research is work intensive, especially in the initiation phase.

Table 6- Allocation of products to CPI-Elementary aggregates - manual

Article group	Article Description	CPI-EA-code	CPI-EA-Description (short)
Diaper Jumbo-pack	Pampers Simp. Dry Gr4 Maxi **7-18kg** 40piece	→no allocation	-
Diaper Jumbo-pack	Pampers BD Gr.3 Midi **4-6kg** GP 136piece	→367	Disposable Diaper, f. 3-6 kg-Baby
Diaper Stock pack	babylove MegaPack Maxi 126piece VL **→Internet research necessary because of missing weight information**	→?	Disposable Diaper, f. 3-6 kg-Baby
Diaper Jumbo-pack	Pampers Simp. Dry Gr5 Jun. **11-25kg** 34Stk	→ no allocation	-
…	…	…	..

There are other options to solve the problem of insufficient article characteristics when mapping scanner data to elementary aggregates and COICOP (sub-) class. EAN-Dictionaries may be obtained from market research companies, hereby upgrading the rudimentary product information in the original scanner data from the retailers with additional product characteristics[4]. Also, retailers might be willing to build up and maintain a harmonized classification system based on COICOP. However, retailers usually want to avoid any additional data management tasks.

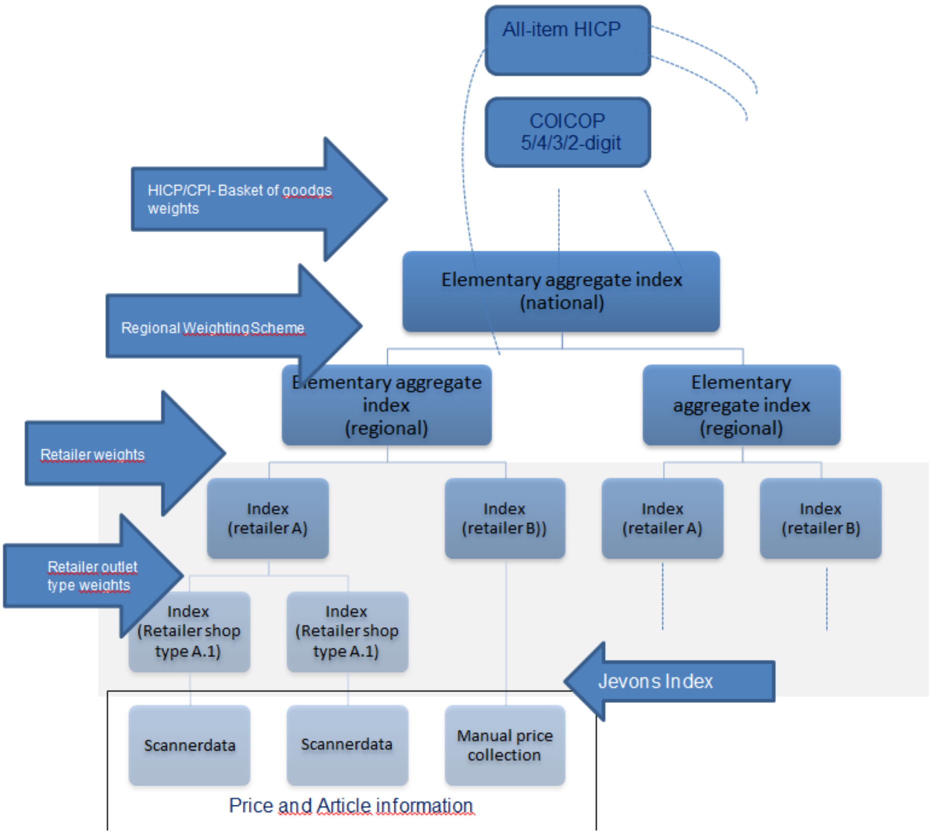
[4] INSEE in France purchases EAN documentation (EAN dictionary) from a market research institute.
Statistic Sweden promotes the development of an extensive international EAN database in collaboration with GS1, the producer of EAN Codes.

ANNEX 2 - AGGREGATION OF ELEMENTARY AGGREGATE INDICES IN THE AUSTRIAN HICP/CPI WITH SCANNER DATA

Scanner data are not fully available for the entire Austrian market. Therefore some changes are needed to integrate scanner data in the existing HICP/CPI compilation process. Currently, price data is not weighted until the level of regions (there are 20 CPI regions in Austria). With scanner data additional levels of aggregation enter the compilation process, namely at retailer level.

Chart 1 below shows how scanner data necessitates the introduction of new strata (retailer and retailer outlet type strata) into the Austrian CPI compilation (grey background). To combine scanner data and classical price collection data, expenditures by retailer and their specific outlet types are used to calculate the weights for these new strata.

Chart 1 - *overall price index compilation for the Austrian HICP with scanner data*


All-item HICP
COICOP 5/4/3/2-digit
HICP/CPI- Basket of goodgs weights
Elementary aggregate index (national)
Regional Weighting Scheme
Elementary aggregate index (regional)
Elementary aggregate index (regional)
Retailer weights
Index (retailer A)
Index (retailer B))
Index (retailer A)
Index (retailer B)
Retailer outlet type weights
Index (Retailer shop type A.1)
Index (Retailer shop type A.1)
Jevons Index
Scannerdata
Scannerdata
Manual price collection
Price and Article information

ANNEX 3 - IMPLEMENTATION OF WEB SCRAPING USING CLICK-AND-POINT SOFTWARE

The Austrian Web-scraping project performs all project tasks using the web-scraping software *import.io*. The main advantage of the tested software is that no advanced programming skills are needed to perform changes to the web-scraping programs in case of website changes.

The success of automatic data collection depends on the ability of the deployed web-scraper to simultaneously improve the data quality while reducing the overall data collection costs. The working hours spent to collect the prices needed to compile the indices can be substantially reduced. In fact, the actual manual price collection should be completely replaced and the quality of the price indices will be higher due to an increased number of measured price quotes. The working time needed for the automatic price collection method consists of various tasks, such as data importing, data cleaning and data checking.

Selection of software

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

- The software provides a high level of usability and can be easily understood by non-IT price statistics staff members.
- The software provides a surface that enables users with basic IT knowledge to develop basic web-scrapers and change the price collection procedures (e.g. in case of website changes).
- The software provides documentation and is adaptable to the internal IT system.
- IT-specialists verified that the software is safe to operate and that it comes along with appropriate licensing, testability and supportability.

Implementation and maintenance of software and supporting IT infrastructure

The IT department installed the selected software. Maintenance procedures to update and test the software regularly and to provide support needs were set up. Automatic web-scraping within Statistics Austria's domain and firewall has been identified as a potential leak. In order to avoid viruses, hackers etc. to infiltrate the Statistics Austria's internal IT-system the web-scraper software operates within a stand-alone system on a separate server. Employees access the software and the scraped data by using a remote server from their PC.

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.

Usage of automatic price collection for various statistics

In order to maximize the output of the investment into automatic data collection on the internet, the actions include 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. Since August 2015, online job offers is web scraped from Austria most important online job portals and tested for its use in Statistics Austria's Job Vacancy Survey.

-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 analyzed 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 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 another web-scraper software whose results could be automatically compared with the results of the first web-scraper.

Major methodological challenges – CPI price data collection using Web-scraping

The irregular maintenance work needed to run the web-scraper 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.

ANNEX 4 - ENCOUNTERED PROBLEMS WHEN USING CLICK-AND-POINT SOFTWARE

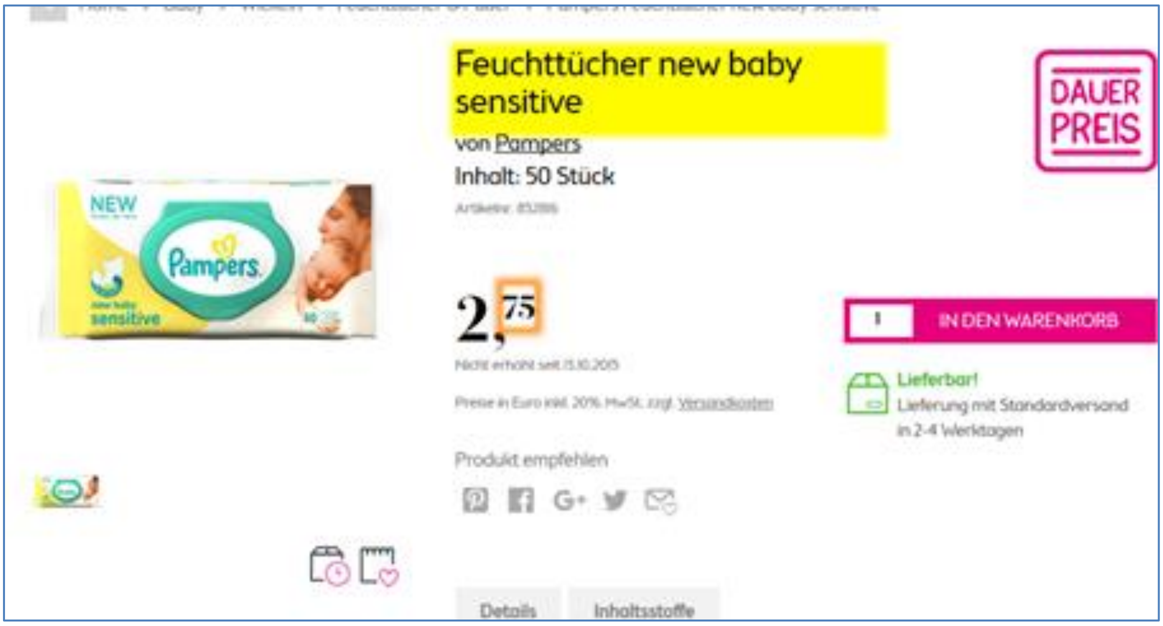
Irregular data fields

When building an extractor, a tailor made solution has to be developed for every website. This is because the data fields on webpages are all individually arranged. In addition to the variety of layouts, websites structures and data fields change regularly. New variables may be added, product hierarchies altered. Typically, the extractor simply extracts non-standardized web data, and standardizing data manipulation is done after exporting the data into the internal price collection infrastructure.

Picture 1 below depicts the product page information of an Austrian drugstore chain. The Price information is split in two fields, EUR and Cents. Therefore, the extractor needs to be trained on two price information fields separately. The returned price data of the extractor will look as follows:

Product Name	(…)	Price EUR	Price CENTS	(…)	(….)
Pampers Feuchttücher new baby sensitive	(…)	2	75	(…)	(…)

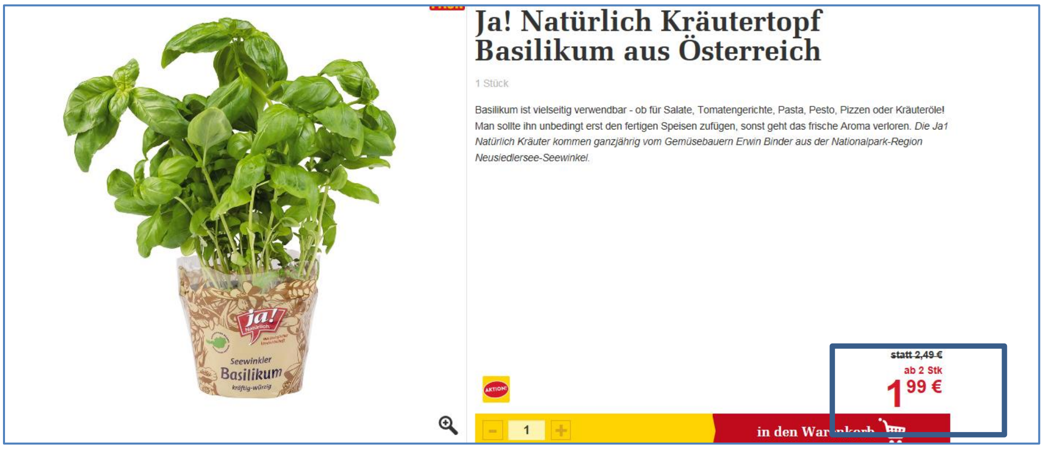
Pic 1 - Irregular price data fields – separate EUR and CENTS data fields


Feuchttücher new baby sensitive
von Pampers
Inhalt: 50 Stück
DAUER PREIS
2,75
IN DEN WARENKORB
Lieferbar!
Lieferung mit Standardversand in 2-4 Werktagen
Preise in Euro inkl. 20% MwSt. zzgl. Versandkosten
Produkt empfehlen
Details
Inhaltsstoffe

Pictures 2 and 3 below depict product page information of an Austrian supermarket. In case of a product promotion, an extractor trained on a normally priced product only, will not return any price data. This is because the position of the price variables of promotion product in HTML Code is different to the position of a normally priced product. Therefore, the extractor needs to be trained on both, price information data fields for normal and promotion prices, respectively. The returned data of the extractor will look as follows:

Product Name	**(…)**	**Price (normal)**	**Price (promotion)**	**Price (before promotion)**	**(….)**
Ja! Natürlich Kräutertopf	(…)	-	1,99	2,49	(…)
Dille Bund aus Österreich	(…)	1,29	-	-	(…)

Pic 2- Irregular price data fields – promotion price

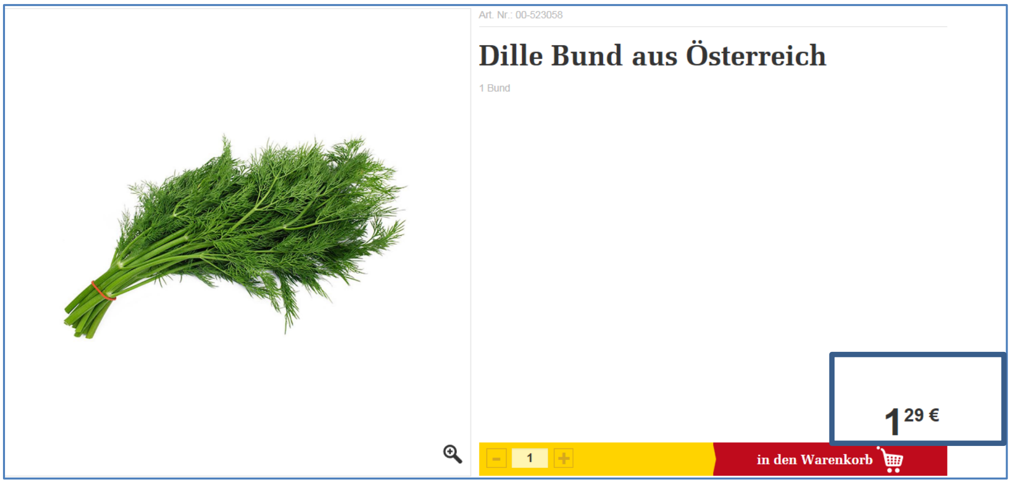


Ja! Natürlich Kräutertopf Basilikum aus Österreich

1 Stück

Basilikum ist vielseitig verwendbar - ob für Salate, Tomatengerichte, Pasta, Pesto, Pizzen oder Kräuteröle! Man sollte ihn unbedingt erst den fertigen Speisen zufügen, sonst geht das frische Aroma verloren. *Die Ja1 Natürlich Kräuter kommen ganzjährig vom Gemüsebauern Erwin Binder aus der Nationalpark-Region Neusiedlersee-Seewinkel.*

Pic 3 - Irregular price data fields – regular price



Art. Nr.: 00-523058

Dille Bund aus Österreich

1 Bund

Irregular data return structures

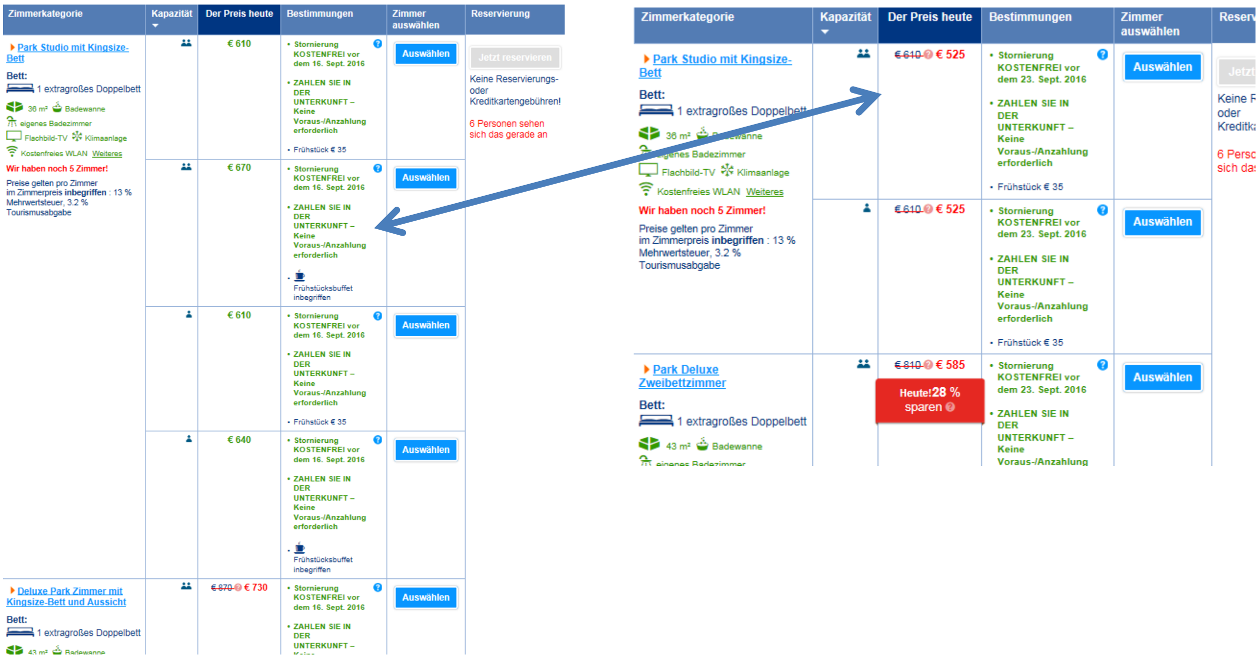
Web scraping has the potential to increase the quality of consumer price indices by improving the coverage of price collection. In particular, price setting developments of accommodation and travel services are more and more volatile. Nevertheless, the complexity of pricing schemes in these industries leads to several practical problems when building price scraping extractors.

Picture 4 below depicts the result page of a specific hotel for a certain room category (studio with king size bed) at two different dates. In order to calculate mean prices for a certain service (e.g. "two people, with breakfast, cancellation possible"), the different sub-categories offered for this need to be scraped and correctly exported. However, the example shows that the web scraper returns different number of results (four and two sub-category results, respectively) depending on the booking date. In the first period with four results, the second line should be selected as it fits the product description best. In the second period, none of the offers technically fit to the CPI product description (breakfast not included). The price for breakfast of 35EUR would have to be added later by CPI staff as the product available.

Workaround:

A web scraper for this hotel websites needs to extract all available price information. Product selection is being done by price statistics staff after the export of all the price data.

Pic 4 - Irregular data returns



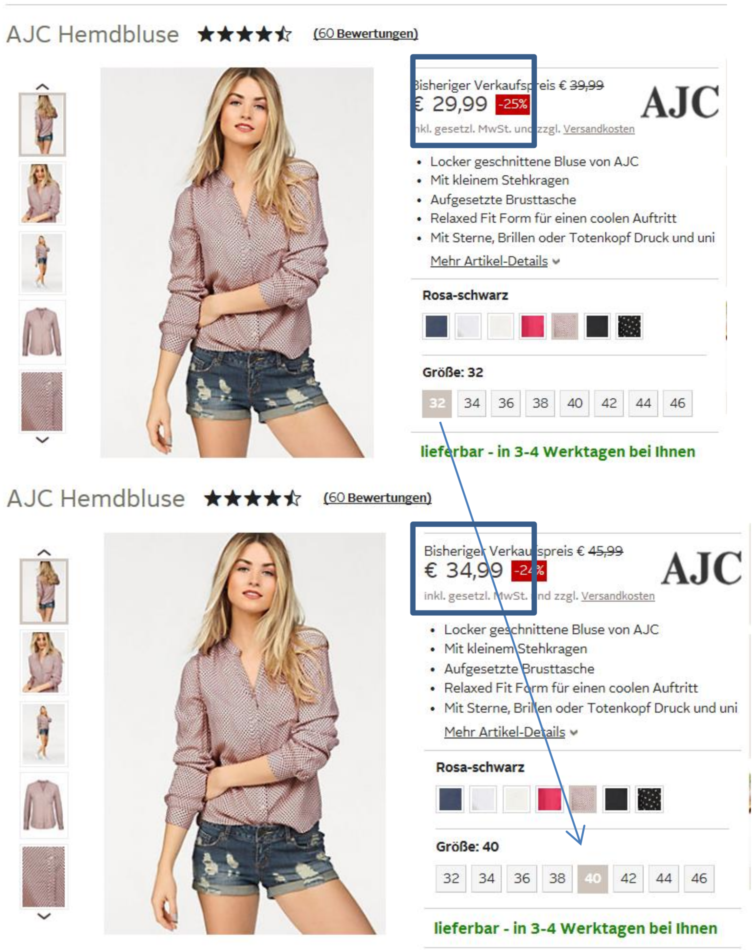
Zimmerkategorie	Kapazität	Der Preis heute	Bestimmungen	Zimmer auswählen	Reservierung
Park Studio mit Kingsize-Bett Bett: 1 extragroßes Doppelbett 36 m² Badewanne eigenes Badezimmer Flachbild-TV Klimaanlage Kostenfreies WLAN Weiteres Wir haben noch 5 Zimmer! Preise gelten pro Zimmer im Zimmerpreis **inbegriffen** : 13 % Mehrwertsteuer, 3.2 % Tourismusabgabe	👥	€ 610	• Stornierung KOSTENFREI vor dem 16. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT – Keine Voraus-/Anzahlung erforderlich • Frühstück € 35	Auswählen	Jetzt reservieren Keine Reservierungs- oder Kreditkartengebühren! 6 Personen sehen sich das gerade an
	👥	€ 670	• Stornierung KOSTENFREI vor dem 16. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT – Keine Voraus-/Anzahlung erforderlich • Frühstücksbuffet inbegriffen	Auswählen	
	👤	€ 610	• Stornierung KOSTENFREI vor dem 16. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT – Keine Voraus-/Anzahlung erforderlich • Frühstück € 35	Auswählen	
	👤	€ 640	• Stornierung KOSTENFREI vor dem 16. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT – Keine Voraus-/Anzahlung erforderlich • Frühstücksbuffet inbegriffen	Auswählen	
Deluxe Park Zimmer mit Kingsize-Bett und Aussicht Bett: 1 extragroßes Doppelbett 43 m² Badewanne	👥	~~€ 870~~ € 730	• Stornierung KOSTENFREI vor dem 16. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT –	Auswählen	

Zimmerkategorie	Kapazität	Der Preis heute	Bestimmungen	Zimmer auswählen	Reserv
Park Studio mit Kingsize-Bett Bett: 1 extragroßes Doppelbett 36 m² Badewanne eigenes Badezimmer Flachbild-TV Klimaanlage Kostenfreies WLAN Weiteres Wir haben noch 5 Zimmer! Preise gelten pro Zimmer im Zimmerpreis **inbegriffen** : 13 % Mehrwertsteuer, 3.2 % Tourismusabgabe	👥	~~€ 610~~ € 525	• Stornierung KOSTENFREI vor dem 23. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT – Keine Voraus-/Anzahlung erforderlich • Frühstück € 35	Auswählen	Jetzt Keine R oder Kreditk 6 Perso sich da
	👤	~~€ 610~~ € 525	• Stornierung KOSTENFREI vor dem 23. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT – Keine Voraus-/Anzahlung erforderlich • Frühstück € 35	Auswählen	
Park Deluxe Zweibettzimmer Bett: 1 extragroßes Doppelbett 43 m² Badewanne eigenes Badezimmer	👥	~~€ 810~~ € 585 Heute!28 % sparen	• Stornierung KOSTENFREI vor dem 23. Sept. 2016 • ZAHLEN SIE IN DER UNTERKUNFT – Keine Voraus-/Anzahlung	Auswählen	

Second level price variation (e.g. according to size)

Web scraping aims at extracting price data to be used for CPI purposes. Price data can easier be integrated in existing index compilation if the regular CPI product descriptions are complied with. For example, Austrian CPI price collection of womens clothing usually collect products with sizes 38 – 42 as these are the most commonly sold sizes. However, clothing web pages usually display prices of the smallest available size first. Prices for other sizes are embedded and can be displayed by clicking on the different sizes categories. Unfortunately, click-and-point web scrapers cannot be trained to select non-standardized buttons (such as the size button as they differ between products). Picture 5 below displays how the price of the shirt changes from 29,99 to 34,99 when selecting size 40, without any changes to the website URL.

Pic 5 - Second level price variation according to size



AJC Hemdbluse ★★★★☆ (60 Bewertungen)

isheriger Verkaufspreis € ~~39,99~~
€ 29,99 -25%
nkl. gesetzl. MwSt. und zzgl. Versandkosten

AJC

- Locker geschnittene Bluse von AJC
- Mit kleinem Stehkragen
- Aufgesetzte Brusttasche
- Relaxed Fit Form für einen coolen Auftritt
- Mit Sterne, Brillen oder Totenkopf Druck und uni

Mehr Artikel-Details

Rosa-schwarz

Größe: 32

32 34 36 38 40 42 44 46

lieferbar - in 3-4 Werktagen bei Ihnen

AJC Hemdbluse ★★★★☆ (60 Bewertungen)

Bisheriger Verkaufspreis € ~~45,99~~
€ 34,99 -24%
inkl. gesetzl. MwSt. und zzgl. Versandkosten

AJC

- Locker geschnittene Bluse von AJC
- Mit kleinem Stehkragen
- Aufgesetzte Brusttasche
- Relaxed Fit Form für einen coolen Auftritt
- Mit Sterne, Brillen oder Totenkopf Druck und uni

Mehr Artikel-Details

Rosa-schwarz

Größe: 40

32 34 36 38 40 42 44 46

lieferbar - in 3-4 Werktagen bei Ihnen

Workaround:

Size selection and product information extraction has to be done at an early stage (e.g. product list). Here, articles can be filtered by size. URL-links change according to set filters. Picture 6 below shows how the URL changes when setting the size filter from “all sizes” to “38, 40 and 42” (arrow 1). The URL without size filter is:

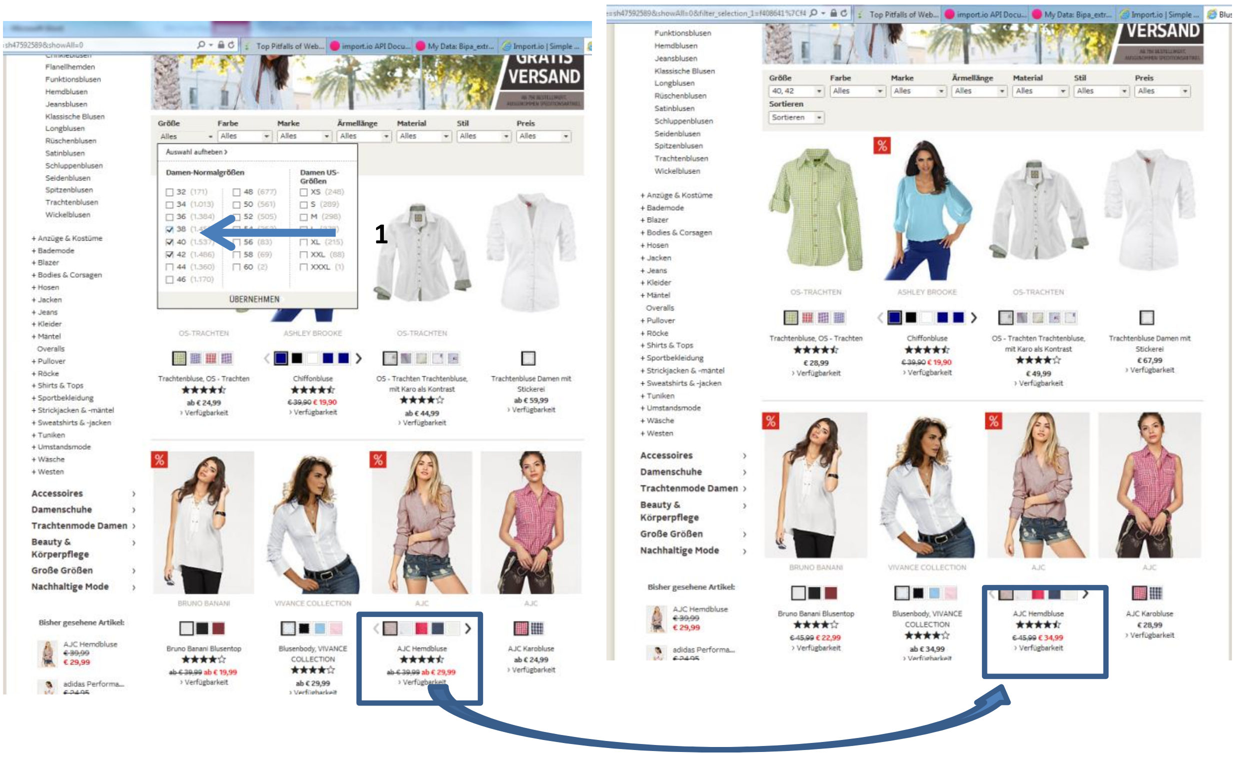
https://www.ottoversand.at/versand/ottoversand-at?CategoryName=sh47592589&showAll=0

The URL with size filter is:

https://www.ottoversand.at/versand/ottoversand-at?CategoryName=sh47592589&showAll=0&filter_selection_1=f408641%7Cf408642

The web scraper needs to be directed at URLs with size filter to extract the price information for the correct sizes as defined in the CPI compilation process.

Pic 6 - Second level price variation according to size – work around



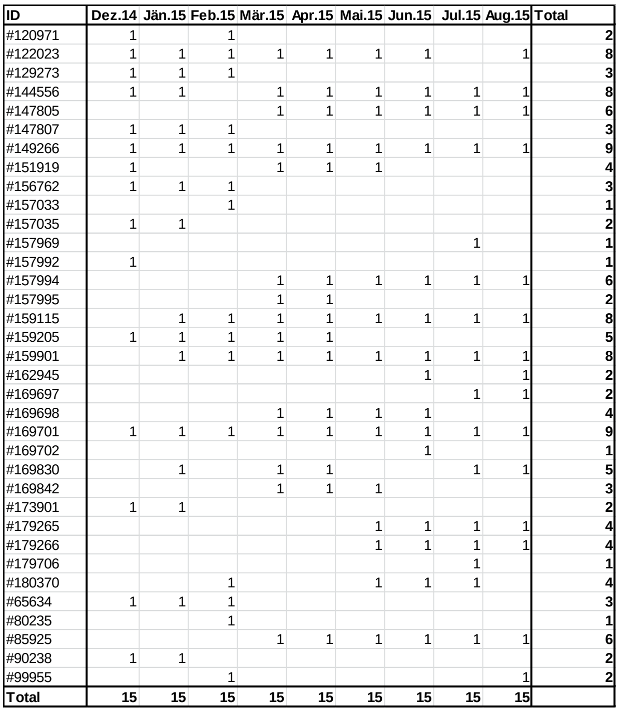
Representative sampling with website data

Price collection with web craping requires sound product sampling procedures. Web scraped data contains no quantity information. In order to allow for a representative sample the scraped price data may be (pre-) filtered according to available quantity information.

Sampling with Pre-Filter

Pre-Filter are often available on websites. Article lists can be ranked by different dimensions such as "best price", "most popular", "best rated" etc. Ranking the articles by "most popular" may be a method to sample well selling articles. However, price collection tests demonstrate that such product rankings might not be as reliable as desired. Table 7 below depicts the product sample of web scraped shampoo product pages. The sample was drawn using the websites' "most popular" sort function. Only two of the sampled products were listed as "most popular" in all of the nine month of the scraping test. Most other "most popular" products appear and disappear from the list. The average listing of a product in the "most popular" list is only about 3.8 month. This may be regarded as an indicator that the "most popular" ranking is used as a marketing tool to promote the launch of certain products. Therefore, the reliability of "most popular" rankings by online retailers should be questioned. Their usage should at least be double checked with the retailer.

Table 7 – Sample of available products featured as "15 most popular" shampoos on retailer website (1=Product price quote available)



ID	Dez.14	Jän.15	Feb.15	Mär.15	Apr.15	Mai.15	Jun.15	Jul.15	Aug.15	Total
#120971	1		1							2
#122023	1	1	1	1	1	1	1		1	8
#129273	1	1	1							3
#144556	1	1		1	1	1	1	1	1	8
#147805				1	1	1	1	1	1	6
#147807	1	1	1							3
#149266	1	1	1	1	1	1	1	1	1	9
#151919	1			1	1	1				4
#156762	1	1	1							3
#157033			1							1
#157035	1	1								2
#157969								1		1
#157992	1									1
#157994				1	1	1	1	1	1	6
#157995				1	1					2
#159115		1	1	1	1	1	1	1	1	8
#159205	1	1	1	1	1					5
#159901		1	1	1	1	1	1	1	1	8
#162945							1		1	2
#169697								1	1	2
#169698				1	1	1	1			4
#169701	1	1	1	1	1	1	1	1	1	9
#169702							1			1
#169830		1		1	1			1	1	5
#169842				1	1	1				3
#173901	1	1								2
#179265						1	1	1	1	4
#179266						1	1	1	1	4
#179706								1		1
#180370			1			1	1	1		4
#65634	1	1	1							3
#80235			1							1
#85925				1	1	1	1	1	1	6
#90238	1	1								2
#99955			1						1	2
Total	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**	

Workaround:

To draw a representative sample, CPI team members filter the available web scraped price data according to available market information. E.g. if the market share of different brands within the cloth segment "jeans" is available/known, team members can filter (and consequently use for index compilation) the web scraped price data for articles of the top 5 brands. This method takes advantage of the knowledge base of CPI price collectors. Often CPI price collectors have years of experience in the respective consumer good segment. A judgmental sample (by brands) based on market knowledge in combination with exhaustive web scraped price data can be regarded as a superior method than dependency of an online retailer ranking and the unweighted usage of every available scraped price information, respectively. The latter includes the danger of using irrelevant price information for the index compilation.

References

ESS quality report.(2014), [Online] Available: http://ec.europa.eu/eurostat/web/quality/quality-reporting

Eurostat (2017). HICP Recommendation on Obtaining Scanner Data (Draft April 2017)

UNECE (2011). Using Administrative and Secondary Sources for Official Statistics. [Online] Available: http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf

Struijs P. and Daas P. (2014). Quality Approaches to Big Data in Official Statistics. Paper presented at the Q2014. [Online] Available: http://www.q2014.at/papers-presentations.html