Blaudow, Christian; Ostermann, Holger

Article

Entwicklung eines generischen Programms für die Nutzung von Web Scraping in der Verbraucherpreisstatistik

WISTA – Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Blaudow, Christian; Ostermann, Holger (2020) : Entwicklung eines generischen Programms für die Nutzung von Web Scraping in der Verbraucherpreisstatistik, WISTA – Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 72, Iss. 5, pp. 103-113

This Version is available at:
https://hdl.handle.net/10419/225308

ENTWICKLUNG EINES GENERISCHEN PROGRAMMS FÜR DIE NUTZUNG VON WEB SCRAPING IN DER VERBRAUCHERPREISSTATISTIK

Christian Blaudow, Holger Ostermann

Schlüsselwörter: Preisstatistik – Verbraucherpreisindex – Web Scraping – Digitalisierung – Scrum

ZUSAMMENFASSUNG

Der Internethandel gewinnt stetig an Verbrauchsbedeutung und ist deshalb in der Berechnung der Inflationsrate entsprechend zu berücksichtigen. Die Messung eines repräsentativen Preises wird zusätzlich durch immer volatilere Preise bei bestimmten Produktgruppen oder Onlinehändlern erschwert, manuell ist die Preismessung bei volatilen Preisen nur noch schwer durchführbar. Aus diesem Grund setzt die Preisstatistik seit Jahren verstärkt auf die Datenerhebung mittels Web Scraping. Um die Verwendung dieser automatisierten Erhebungstechnik zu erleichtern, wurde eine generisch aufgebaute, web-basierte Anwendung entwickelt. Die Anwendung wird derzeit in der Preisstatistik schrittweise eingeführt und soll manuelle Internetpreiserhebungen für den Verbraucherpreisindex und den harmonisierten Verbraucherpreisindex bis Ende 2021 ersetzen.

Keywords: *price statistics – consumer price index – web scraping – digitalisation – scrum*

ABSTRACT

As e-commerce is continuously gaining in importance, it needs to be considered appropriately in calculating the inflation rate. Measuring representative prices has become more difficult because the prices of certain product groups or those specified by online retailers have become increasingly volatile, and a manual measurement of volatile prices is hardly feasible. For this reason, web scraping has been increasingly used to collect data for price statistics in recent years. A generic web-based application has been developed to facilitate the use of this automated data collection technique. The current gradual introduction of this application in price statistics is aimed at replacing the manual internet price collection for purposes of the consumer price index and the harmonised index of consumer prices by the end of 2021.

Christian Blaudow

hat International Economics an der Georg-August-Universität Göttingen studiert und ist Referent im Referat „Verbraucherpreise“ des Statistischen Bundesamtes. Schwerpunkte seiner Arbeit sind die Automatisierung von Preiserhebungen im Internet sowie die Entwicklung von Methoden für die Integration und den Umgang mit großen Datenmengen.

Holger Ostermann

hat Wirtschaftsingenieurwesen an der Fachhochschule Südwestfalen und Computer Science an der FernUniversität in Hagen studiert. Er arbeitet als Referent im Referat „IT-Kompetenzzentrum Datenerhebung“ des Statistischen Bundesamtes unter anderem an Themen wie Web Scraping und den Dateneingängen für den Zensus.

1

Einleitung

In Deutschland kaufen immer mehr Menschen Waren im Internet ein (Statistisches Bundesamt, 2020). Insbesondere während der Corona-Pandemie haben die Einkäufe im Internethandel stark zugenommen (Fentem, 2020; Hansen, 2020b). Große und seit der Corona-Pandemie auch kleinere Einzelhandelsunternehmen verlagern zunehmend den Verkauf ihrer Produkte von stationären Geschäften in ihre Internetshops. Der steigende Anteil der Waren, die im Internet gehandelt werden, hat zur Folge, dass die Stichprobe der Internetpreiserhebung weiter ausgeweitet werden muss.

Die Preiserhebung im Internethandel für den Verbraucherpreisindex (VPI) und den Harmonisierten Verbraucherpreisindex (HVPI) unterliegt, bedingt durch die Digitalisierung, einem großen Wandel. Zum einen muss die Erhebung der steigenden Verbrauchsbedeutung des Internethandels Rechnung tragen, indem mehr online gehandelte Produkte und mehr online handelnde Unternehmen in die Stichprobe aufgenommen werden. Der Anteil der Artikel, für die Preise im Internet erhoben werden, hat sich im VPI beziehungsweise HVPI seit dem Basisjahr 2005 bis zum Basisjahr 2015 von 5 % auf mehr als 10 % erhöht. Dies gilt zumindest bei den Gütern des Warenkorbs, für die eine Einteilung der Konsumausgaben der privaten Haushalte in verschiedene Geschäftstypen möglich ist.[1]

Zum anderen ist für die Messung eines repräsentativen Preises teilweise die Frequenz der Datenerhebung zu erhöhen. Derzeit werden für die Verbraucherpreisstatistik jeden Monat etwa 10 000 Preise für Waren und Dienstleistungen im Internet noch zu einem großen Teil manuell erhoben. Studien über das Preissetzungsverhalten von Onlinehändlern haben gezeigt, dass sich Preise im Internethandel mehrfach im Monat ändern (Blaudow/Burg, 2018; Hansen, 2020a). Eine höhere Frequenz dieser Preiserhebungen ist durch einen verstärkten Einsatz automatisierter Verfahren wie Web Scraping[2] ohne großen Zusatzaufwand zu erreichen.

1 Hierbei muss festgehalten werden, dass sich die Geschäftstypengewichtung nur auf etwa ein Drittel aller Güter im Warenkorb bezieht (Sandhop, 2012).

2 Web Scraping (englisch: scraping = kratzen/abschürfen) bezeichnet das automatisierte Auslesen von Daten auf Internetseiten.

Der Artikel beschreibt im folgenden Kapitel 2 den bisherigen Einsatz von Web Scraping in der Verbraucherpreisstatistik. Dieser hat zur Idee geführt, ein generisches Programm für das Web Scraping in der Verbraucherpreisstatistik zu entwickeln (Kapitel 3). Kapitel 4 erläutert die Herausforderungen des Einsatzes von Web Scraping in der Datenerhebung, der Prozess einer Datenerhebung mithilfe des generischen Programms wird in Kapitel 5 dargestellt. Die neuen Validierungsprozesse in der Verbraucherpreisstatistik schildert Kapitel 6; der Beitrag schließt mit einem Fazit und Ausblick in Kapitel 7.

2

Der Einsatz von Web Scraping in der Verbraucherpreisstatistik

Die manuelle Datenerhebung im Internet ist zeitlich sehr aufwendig und für die verfügbare Bearbeitungszeit einer Monatsstatistik optimiert. Durch den Einsatz von Web Scraping ist es möglich, bei gleichbleibendem Aufwand die Datenerhebung auszuweiten. Diese Technik besitzt das Potenzial, die Datenerhebung im Internet weitgehend zu automatisieren. Dabei werden zu bestimmten Zeitpunkten automatisch zuvor definierte Preise und andere Informationen eines Artikels im Internet extrahiert. Preiserhebungen können zu jedem beliebigen Zeitpunkt beginnen und fast unendlich häufig wiederholt werden, ohne von Arbeitszeiten oder von einer in gleichem Maße erhöhten Personalverfügbarkeit abhängig zu sein. Web Scraping ermöglicht mit beschränktem Zusatzaufwand, größere Datenmengen zu erheben und somit die Stichprobe der Internetpreiserhebung zu erweitern.

Insgesamt profitiert die amtliche Statistik von Web Scraping in großem Umfang, weil es weitgehend ohne manuellen Einsatz die Erfassung von deutlich mehr Preisen aus dem Internet ermöglicht. Dies ist insbesondere in solchen Produktbereichen beziehungsweise bei solchen Onlineshops hilfreich, bei denen die Preise sehr volatil sind.

Darüber hinaus können Datenerhebungen im Internet die Belastung der Auskunftspflichtigen senken. Durch die zunehmende Verbreitung des Internethandels können – eine einheitliche Preisgestaltung der entspre-

chenden Anbieter vorausgesetzt – Erhebungen auf Webseiten Erhebungen in örtlichen Filialen teilweise oder auch vollständig ersetzen. Da diese Erhebungen „geräuschlos“ ablaufen und bei Verwendung einer zuvor definierten Stichprobe mit weitgehend sequenziellem Ablauf die Funktionsfähigkeit der Webseiten nicht spürbar beeinträchtigen, werden weder das Verkaufspersonal noch die Verwaltung belastet.

Seit dem Jahr 2012 untersucht das Statistische Bundesamt die Möglichkeiten automatisierter Preiserhebungen im Internet mithilfe von Web Scraping (Brunner, 2014). Finanziert wurden diese Arbeiten weitgehend durch das Statistische Amt der Europäischen Union (Eurostat). Dabei wurde von Beginn an ein imitierender Ansatz gewählt, bei dem einzelne, manuelle Preiserhebungen im Internet automatisiert werden. Das heißt, zu bestimmten Zeitpunkten werden automatisiert Preise für zuvor definierte Produkte extrahiert, die ansonsten manuell erhoben wurden. Die Internetseite, auf der das jeweilige Produkt angeboten wird, kann häufig mit der entsprechenden URL[3] direkt aufgerufen werden. Ein anderer Weg ist, anhand von Informationen, die in einer Datenbank liegen oder im Programm errechnet werden, Schritt für Schritt zur Internetseite zu steuern. Das Programm erkennt anhand der festgelegten Definition das Produkt und extrahiert den gesuchten Preis. Die so imitierte manuelle Arbeit der Preiserheberinnen und Preiserheber kann in fast beliebiger Frequenz wiederholt werden. Im Gegensatz zu diesem Ansatz mit gezielter Produktsuche, dem sogenannten targeted scraping, könnten beispielsweise die Preise aller verfügbaren Produkte eines Onlinehändlers ohne vorausgehende manuelle Produktauswahl abgegriffen werden (sogenanntes bulk scraping). Diesen Ansatz verfolgt das Statistische Bundesamt bisher nicht, jedoch sollen in einer zeitnahen Studie die Vor- und Nachteile dieses Verfahrens evaluiert werden.

3

Die Entwicklung des generischen Programms „ScraT“

3.1 Die Idee eines generischen Programms

Bisher wurde Web Scraping in der Preisstatistik isoliert vom hausinternen Netzwerk des Statistischen Bundesamtes mit projekteigener IT-Ausstattung (nicht serverbasiert) und separater Internetverbindung betrieben. Dies führte zu hoher Abhängigkeit von einzelnen Geräten und barg die Gefahr von Systemabstürzen. Die Programmierung konnte ebenso wie die Programm- und Datenpflege nur von wenigen Personen mit speziellen IT-Kenntnissen durchgeführt werden.[4] Das schuf hohe Abhängigkeiten von einzelnen Personen und Programmen. Die Programme wurden zwar vereinheitlicht und Beschäftigte an Scrapingtools geschult, jedoch reichten diese Maßnahmen nicht aus, um die manuellen Preiserhebungen vollständig, effizient und produktionssicher zu ersetzen.

Alle Web-Scraping-Programme in der Preisstatistik sind in der Programmiersprache Java geschrieben, die auch die Realisierung von grafischen Benutzeroberflächen und Webanwendungen unterstützt. Die Kombination aus der Nutzung von grafischen Benutzeroberflächen mit vereinheitlichten Programmcodes und dem Wunsch, Web Scraping auch ohne fundierte IT-Kenntnisse anzuwenden, führte zur Idee, ein generisches Programm mit einfach zu bedienenden grafischen Oberflächen zu entwickeln. Bei einem generischen Programm werden Funktionen und Methoden zunächst sehr allgemein entworfen, um sie für alle möglichen Datentypen und -strukturen verwenden zu können. Je nach Anforderung einer Erhebung werden spezielle Funktionen nach und nach hinzugefügt. Der Aufbau einer Erhebung hat somit großen Wiedererkennungswert für Personen, die die Programme nutzen, warten und anpassen müssen. Im Kern sind alle Programme gleich aufgebaut, nur die Spezialisierungen sind unterschiedlich. Im Idealfall können Programme lediglich durch „Zusammenklicken“ von Spezialisierungen erstellt werden.

3 Uniform Resource Locator.

4 Für zusätzliche Erläuterungen siehe Blaudow (2018).

Mit der Einführung eines generischen Programms können zum einen die bestehenden Scrapingprogramme vereinheitlicht werden, indem sie in das generische Programm übertragen werden. Zum anderen kann eine Vielzahl an neuen automatisierten Erhebungsprogrammen erstellt werden, welche derzeit noch manuell ablaufen. Durch den gemeinsamen Kern ist auch die technische Wartung jeder Erhebung personenunabhängig.

3.2 Die agile Entwicklungsphase

Im Oktober 2018 hat die Entwicklung des generischen Programms für die nutzerfreundliche Anwendung von Web-Scraping-Techniken in der Preisstatistik begonnen.

Die agile Entwicklung erfolgte intern im Statistischen Bundesamt, basierend auf kostenlosen Open-Source-Komponenten und mit Scrum (Schwaber/Sutherland, 2017) als leichtgewichtigem Vorgehensmodell. Das Produkt erhielt nach kurzer Zeit die Bezeichnung „ScraT“, ein Kompositum aus den Wörtern „Scraping“ und „Tool“.

Das Scrumteam bestand aus dem Product Owner, dem Scrum Master und dem Entwicklungsteam. Gemäß Scrum wurden die Anforderungen während der gesamten Projektlaufzeit vom Product Owner in einer Liste, dem Product Backlog, gepflegt und priorisiert. Außerdem wurden User Stories (Plewa, 2019) erstellt, die aus Anwendersicht jeweils eine (Teil-)Funktion des gewünschten Systems einfach beschreiben und begründen. Bei den in definierten Zeitabständen stattfindenden Sprint Planning Meetings des gesamten Teams wurden dann aus diesem Product Backlog die Anforderungen für die Umsetzung im nächsten Sprint entnommen. Das Team teilte dieses Arbeitspaket (Increment) in kleine Aufgaben (Tasks) auf, die im Sprint Backlog beziehungsweise Scrum Board mit den bearbeitenden Personen festgehalten wurden. Während der Umsetzung fanden werktägliche Daily Scrum Meetings des Entwicklungsteams mit dem Scrum Master statt. Am Ende eines Sprints erfolgte das Sprint Review Meeting, bei dem die Umsetzung des Sprints mit dem Product Owner besprochen wurde. Zusätzlich fand in der Regel noch eine kurze Sprint-Retrospektive zum Analysieren des Sprints statt. Falls noch nicht umgesetzte (Teil-)Aufgaben existierten, wurden diese wieder in das Backlog für das dann folgende Sprint Planning Meeting aufgenommen.

3.3 Die Eigenschaften von ScraT

Den IT-Standards der Statistischen Ämter des Bundes und der Länder folgend werden Java 8 und JSF (Java Server Faces) mit PrimeFaces für die Web-Anwendungsentwicklung verwendet und Tomcat als Webserver eingesetzt. Als Datenbank wird MySQL ausgewählt und über JPA (Java Persistence API) mit Hibernate angebunden. Für die Ablaufsteuerung der eingeplanten Web-Scraping-Aufgaben wurde Quartz verwendet. ScraT soll den manuellen Preiserhebungsvorgang automatisieren und dabei auch mehrere Schritte auf Webseiten ausführen, bei denen Daten interaktionsabhängig durch JavaScript nachgeladen werden. Dies können Web Browser am besten gewährleisten und mit Selenium[5] gibt es ein Softwarepaket, das es ermöglicht, mit Selenium WebDriver gemäß der W3C WebDriver-Spezifikation[6] Browser zu steuern. Das Selenium Framework unterstützt momentan mit Chrome, Firefox, Edge, Internet Explorer, Opera und Safari alle aktuell (im August 2020) gängigen Browser und die Browserhersteller stellen entsprechende Treiber zur Verfügung.[7] Einen großen Vorteil stellt hierbei die Verwendung des gleichen Codes für die Steuerung der Abläufe in unterschiedlichen Browsern dar, sodass nur noch eine browserspezifische Einbindung von Browsertreiber und Installationspfad benötigt wird. Selenium bietet Unterstützung für die Programmiersprachen Ruby, Java, Python, C# und JavaScript.[8] Daher wird Selenium mit Chrome und Firefox als installierten Browsern verwendet. Diese zwei Browser unterstützen über Erweiterungen auch die Ermittlung eines Elements einer Webseite per Mausklick. Der gesamte Erhebungsprozess einschließlich der Navigation im Internet wird mit dem festgelegten Internetbrowser durchgeführt, um Ladezeiten und Skriptfehler zu minimieren.

Das generische Programm wurde als nutzerfreundliche Anwendung konzipiert: Die Benutzerinnen und Benutzer müssen keine Codes programmieren, sondern lediglich Abläufe anlegen, Variablen definieren und HTML-Posi-

5 Informationen zu Selenium bietet die offizielle Selenium-Webseite unter www.selenium.dev

6 Für weitere Information zur W3C WebDriver-Spezifikation siehe Stewart/Burns (2020).

7 Die Browserunterstützung von Selenium wird auf folgender Seite beschrieben: www.selenium.dev

8 Für weitere Informationen zu Seleniums Programmiersprachenunterstützung und Download siehe die offizielle Selenium-Downloadseite unter www.selenium.dev/downloads

tionen der Informationen auf Internetseiten in entsprechenden Masken einfügen. Die für die Identifizierung der zu erhebenden Produkte notwendigen Eingangsdaten werden außerhalb der Anwendung gepflegt und aktualisiert. Für die Erhebung können diese Eingangsdaten dann im CSV[9]-Format in die Anwendung importiert werden. Die Anwendung ist in der Lage, erhobene Werte in zusätzlichen Variablen für die weitere Verarbeitung im Programm anzulegen. Die Ausgabedaten, also die Ergebnisse der automatisierten Preiserhebungen, werden in einer Datenbank abgelegt; sie können in den Formaten CSV, XML[10] oder XLSX[11] betrachtet und exportiert werden.

Die automatisierten (Web Scraping-)Erhebungen (sogenannte Workflows) sollen möglichst benutzerfreundlich erstellt werden. Daher wurde ein einfaches und flexibles Modell für diese Workflows entwickelt, die beispielsweise aus verschachtelten einzelnen Schritten mit bestimmten Aktionen wie URL eingeben, Klick oder Lesen eines Werts bestehen. Einen solchen Workflow-Ausschnitt in der Web-Oberfläche von ScraT zeigt ↘ Grafik 1.

Ein Workflow wird dann in einem dafür entwickelten XML-Format komprimiert abgespeichert. ↘ Grafik 2 stellt einen vereinfachten Schritt mit einer Leseaktion dar.

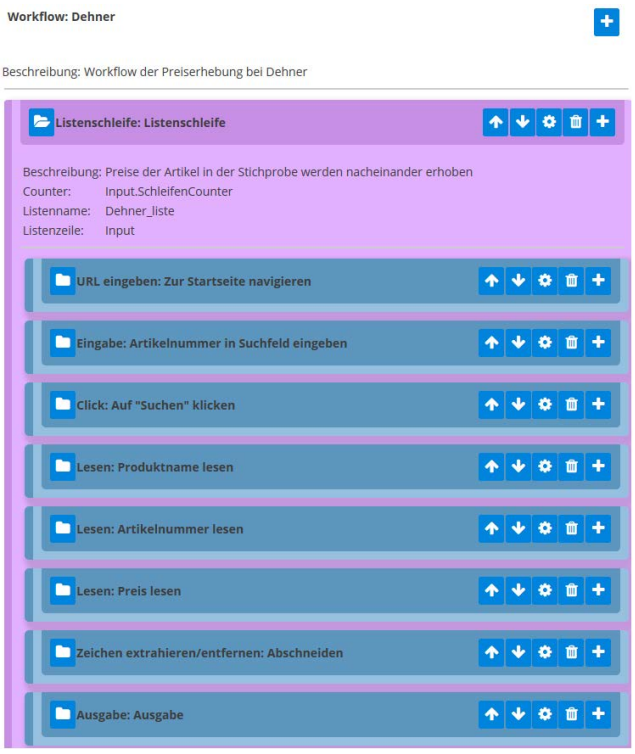
9 Comma Separated Values.

10 Extensible Markup Language.

11 Excel Spreadsheet von Microsoft Excel 2007 oder höher, das auf XML basiert.

Grafik 1

Darstellung eines Workflow-Ausschnitts über die Web-Oberfläche von ScraT



2020 - 0496

Des Weiteren läuft ScraT serverbasiert in einer sicheren Linux-Produktionsumgebung des Informationstechnikzentrums Bund (ITZBund). Damit ist die Abhängigkeit von einzelnen Personen, isolierten Geräten und separater Internetverbindung nicht mehr gegeben.

Grafik 2

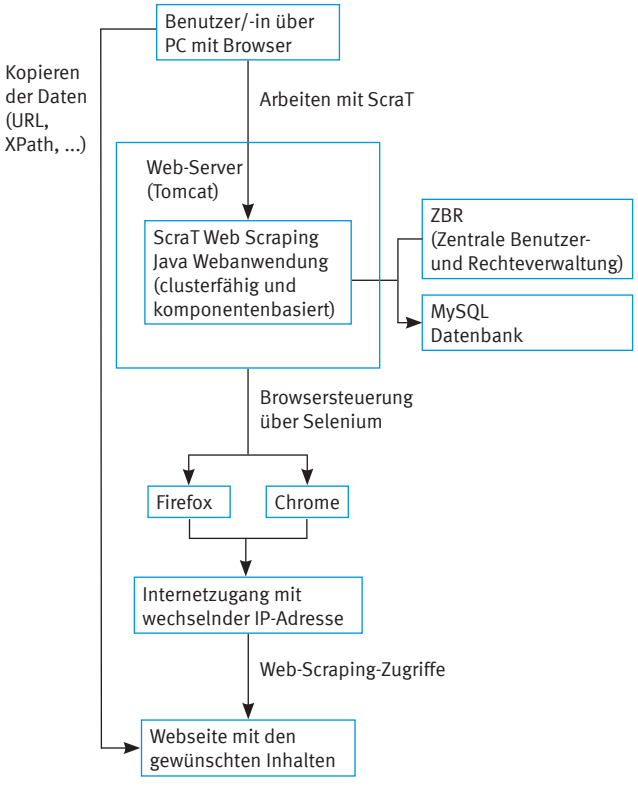
Vereinfachter Schritt eines Workflows mit einer Leseaktion im XML-Format

```
<step _class="de.destatis.webscraping.workflow.action.ReadAction">
  <uuid>f65caead-e156-4f16-ab8b-9f0799be518e</uuid>
  <name>Artikelnummer lesen</name>
  <description></description>
  <parent>8d57e926-2c9e-473c-9159-187047dd17b9</parent>
  <maxRetries>2</maxRetries>
  <xPath _class="de.destatis.webscraping.workflow.information.InformationTypeString">
    <name></name>
    <value>.//*[@id='articleNr']/span</value>
    <rawValue>.//*[@id='articleNr']/span</rawValue>
    <contentType>CONSTANT</contentType>
  </xPath>
  <outputVariable _class="de.destatis.webscraping.workflow.information.InformationTypeString">
    <name>Artikelnummer</name>
    <value/>
    <rawValue></rawValue>
    <contentType>VARIABLE</contentType>
  </outputVariable>
  <timeoutReadStepMilliSeconds>-1</timeoutReadStepMilliSeconds>
</step>
```

2020 - 0497

↘ Grafik 3 stellt die entwickelte Lösung dar.

Grafik 3
Diagrammdarstellung der entwickelten Lösung ScraT mit Interaktion der Benutzerin oder des Benutzers und Web-Scraping-Zugriffen



Eine zentrale Benutzer- und Rechteverwaltung (ZBR) wurde angebunden. Um eine Zugriffssperre (Blacklisting) aufgrund zu vieler Zugriffe in zu kurzer Zeit zu vermeiden, wird vom ITZBund ein Internetzugang mit wechselnder IP-Adresse über eine Proxyserver-Farm bereitgestellt. Aus einem 27-Bit-Netz für wechselnde IP-Adressen wird abwechselnd jede IP-Adresse verwendet (Round-Robin-Verfahren).

Der Benutzer oder die Benutzerin arbeitet mit zwei Browserfenstern parallel. Das eine Browserfenster wird für die Webseite mit den gewünschten Inhalten verwendet, damit von dort die benötigten Informationen, wie URL und Positionen der relevanten Elemente, kopiert werden können. Im zweiten Browserfenster wird der Web-Scraping-Workflow bearbeitet und hier werden die kopierten Informationen eingetragen. Zusätzlich werden bei Bedarf CSV-Daten importiert; anschließend erfolgt die Zeitplanung der Erhebungen. Soll eine Erhebung durchgeführt werden, wird der vorgesehene Browser für diese Erhebung instanziiert und die Daten werden mittels des zuvor definierten Web-Scraping-Workflows ermittelt und in der Datenbank gespeichert. Anschließend können sie über ScraT in unterschiedlichen Formaten (zum Beispiel CSV) heruntergeladen werden. Neben der Installation auf einem Server beziehungsweise Servercluster mit gemeinsamer Datenbank und ZBR-Anbindung kann die entwickelte Lösung für Testzwecke auch eigenständig portabel ohne Installation von einem definierten Laufwerkpfad eines Arbeitsplatz-PCs verwendet werden.

4
Herausforderungen beim Web Scraping

Geplant ist, mit dem generischen Web-Scraping-Programm sukzessive die bisherigen manuellen Preiserhebungen im Internet abzulösen. Damit können in kurzer Zeit größere Datenmengen erhoben werden und die Datenerhebungen werden flexibler. So können beispielsweise Erhebungen an Wochenenden oder ohne großen Aufwand Nacherhebungen durchgeführt werden. Allerdings kann die Wartungsintensität noch hoch ausfallen, wenn eine Internetseite häufig verändert wird (technische Validierung) oder die zu erhebenden Produkte häufig wechseln und ersetzt werden müssen (fachliche Datenvalidierung, siehe Abschnitt 6.3).

Unter technischer Validierung wird die Anpassung des Workflows in ScraT bezeichnet. Typische Ursachen für Anpassungen stellen beispielsweise Positionsänderungen von Elementen auf einer Webseite dar, die aus Layoutänderungen resultieren. Der komplette Workflow einer Preiserhebung muss dafür nicht verändert werden, sondern lediglich eine Zeile in einem Schritt des Programms. Schritte im Workflow, bei denen ein Wert eingetragen, ein Klick getätigt oder eine Information extrahiert werden muss, sind immer mit einer Position auf der jeweiligen Seite versehen. ScraT nutzt die Abfragesprache XPath[12], um ein Element auf einer Seite im HTML5(HyperText Markup Language Version 5[13])-For-

12 Für weitere Informationen zu XPath (XML Path Language) siehe die XPath W3C Standards und Entwürfe; verfügbar unter www.w3.org/TR/xpath

13 Für weitere Informationen zu HTML5 siehe die HTML Living Standard Spezifikation, verfügbar unter html.spec.whatwg.org

mat zu adressieren. Je nach Qualität des Seitenquelltexts kann die technische Wartung häufig oder selten nötig sein. Prinzipiell gilt, dass Seitenquelltexte mit eindeutigen IDs bei den Elementen der zu extrahierenden Daten für Web Scraping in der Regel stabile Seiten darstellen. Sie erfordern somit wenig Wartungsaufwand. Ideal sind Informationen eines Elements, welches im Seitenquelltext direkt durch eine ID adressiert werden kann. Ein technisches Grundverständnis über den Aufbau von Internetseiten ist notwendig. Spezielle Schulungen werden die Einführung des generischen Programms begleiten.

Stabiler XPath:

.//*[@id='priceblock_ourprice']

Sofern die eindeutige Adressierung über die ID des sichtbaren Elements konstant bleibt, kann das Layout der Webseite beliebig geändert werden, ohne dass das Web-Scraping-Programm angepasst werden muss.

Unsicherer XPath:

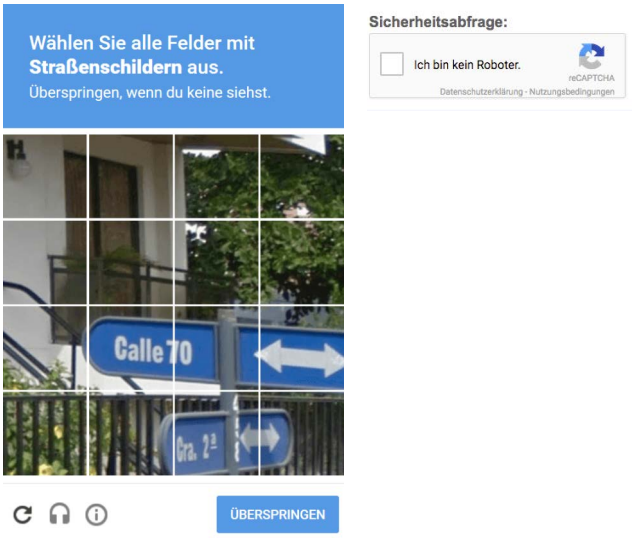
.//*[@id='prodDetails']/div/div[2]/div[1]/div[2]/div/div/table/tbody/tr[1]/td[2]

Hier verfügt nur ein Element auf einem deutlich höheren Level über eine eindeutige ID, sodass ein relativ langer Pfad angegeben werden muss. Falls eine Layoutänderung der Webseite Auswirkungen auf den Pfad hat, muss der XPath für das Web Scraping auch entsprechend angepasst werden.

ScraT ist für alle Datenerhebungen im Internet nutzbar, bei denen die Informationen frei verfügbar sind und nicht in Bildern oder Dokumenten vorliegen. Bei der regulären Datenerhebung im Internet können darüber hinaus weitere Probleme auftreten. Onlinehändler schützen sich vor Zugriffen von übermäßig vielen Robotern, indem sie nach einer bestimmten Anzahl oder Frequenz von Zugriffen oder Klicks einen Captcha[14] einblenden. ↘ Grafik 4

ScraT und vermutlich alle vergleichbaren Roboter sind Stand August 2020 ohne zusätzliche KI(Künstliche Intelligenz)-Erweiterungen nicht dazu in der Lage, die Captcha-Denkaufgaben zu lösen oder das System hinter der Sicherheitsabfrage zu überlisten. Aus diesem Grund ist es wichtig, häufiger und an verschiedenen Tagen Preiserhebungen mit limitiertem Umfang und limitierter Geschwindigkeit durchzuführen, um die Gefahr zu verringern, dass die Erhebung in einem Captcha endet. Außerdem reduziert der vom ITZBund für ScraT bereitgestellte Internetzugang mit dynamisch wechselnden IP-Adressen das Risiko einer Sperrung durch den Webseitenanbieter deutlich. Zusätzlich bietet es sich an, beispielsweise wöchentlich und automatisiert in die erhobenen Daten zu schauen, um die Vollständigkeit zu überprüfen.

Grafik 4
Beispiel für einen Captcha



2020 - 0499

5 Prozess der Datenerhebung mithilfe von ScraT

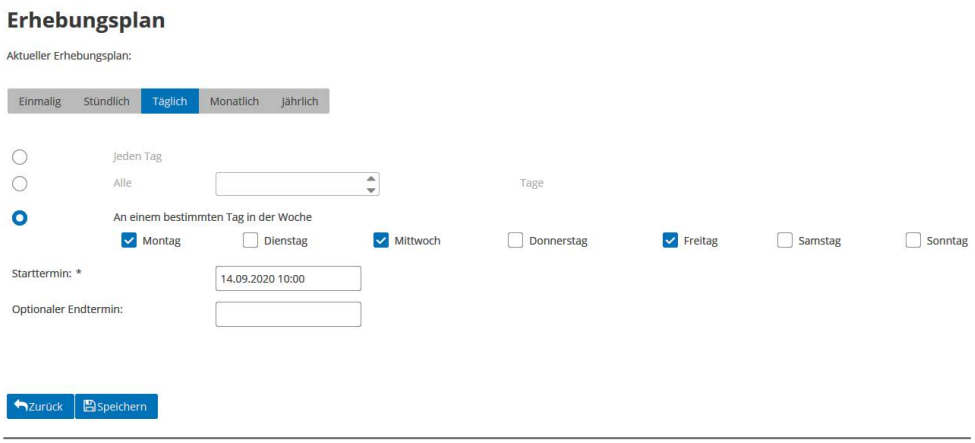
Mit dem Einsatz des generischen Programms ist der Prozess der Datenerhebung im Vergleich zum bisherigen Vorgehen zu erweitern:

Zu Beginn sind die wesentlichen Informationen zum Produkt, wie Name, Artikelnummer und exakte Internetadresse (URL), in einer Tabelle zusammenzufassen. Hinzu kommen wichtige Informationen für die spätere Weiterverwendung, wie die Kennzeichnungen für Güter (COICOP[15]), Geschäfte (Berichtsstellen) und Produkt-

14 Captcha: Test zur Unterscheidung von Computern und Menschen; englisch: completely automated public Turing test to tell computers and humans apart.

15 COICOP: Classification of Individual Consumption by Purpose – Klassifikation der Verwendungszwecke des Individualverbrauchs.

Grafik 5
Beispiel eines Zeitplans einer Erhebung, welche immer montags, mittwochs und freitags um 10 Uhr beginnt



2020 - 0500

varianten. Diese auch bisher schon erforderlichen Eingangsdaten werden in die Anwendung importiert und für die weitere Verwendung als Variablen angelegt. Daraufhin kann die Benutzerin oder der Benutzer den Ablauf der Preiserhebung definieren. So besteht die Möglichkeit, das Programm direkt zur Produktseite navigieren zu lassen oder schrittweise durch mehrere Seiten zu steuern, beispielsweise über Filter. Alternativ können die Datenerhebungen mit Schleifen und Bedingungen präziser ablaufen und die Qualität der Ergebnisse gesteigert werden. In einem Zeitplan kann der Benutzer oder die Benutzerin festlegen, ob die Erhebung stündlich, täglich, wöchentlich oder jährlich wiederholt werden soll. Es ist ebenfalls möglich, die Tage und genauen Uhrzeiten einzustellen. ↘ Grafik 5

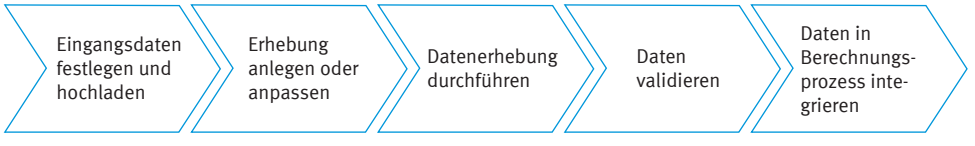
Unmittelbar nach einer Datenerhebung oder spätestens zum Ende eines gesamten Berichtszeitraums validieren die zuständigen Preiserheberinnen und Preiserheber die erhobenen Daten. Dieser Prozess ist in die im nachfolgenden Kapitel näher beschriebenen vier Schritte unterteilt. Nach der Validierung werden die plausiblen Preise entweder direkt in das Berechnungsprogramm der Verbraucherpreisstatistik übernommen oder es müssen Preisindizes berechnet werden, welche als sogenannte Messzahlen in den weiteren Produktionsprozess einfließen. Damit wäre ein Erhebungsprozess abgeschlossen und kann wieder neu starten. ↘ Grafik 6

6

Neue Validierungsprozesse in der Verbraucherpreisstatistik

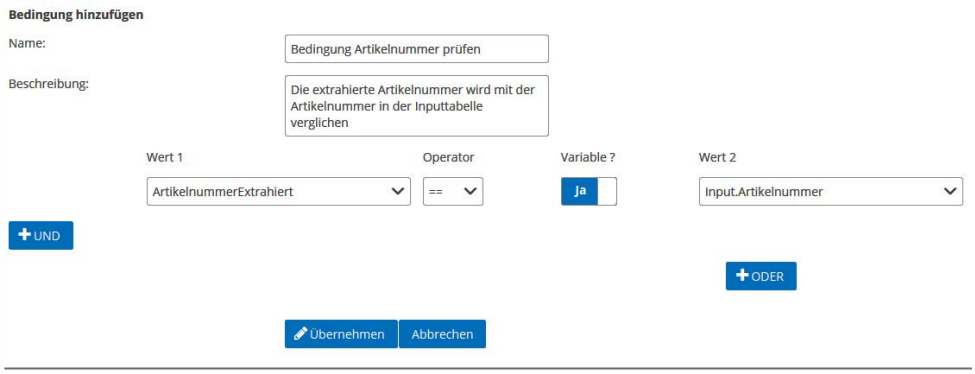
Die Verwendung großer Datenmengen in der monatlichen Statistikproduktion setzt einen umfassenden und effizienten Validierungsprozess voraus. Insbesondere bei der Preiserhebung im Internet können unplausible Ergebnisse das Gesamtergebnis verfälschen. Aus diesem Grund müssen dem aktuellen Validierungsprozess in der Verbraucherpreisstatistik, der auf eine manuelle Preiserhebung ausgerichtet ist, weitere Validierungsprozesse vorgeschaltet werden.

Grafik 6
Prozess der Datenerhebung mithilfe des generischen Programms



2020 - 0501

Grafik 7
Beispiel für die Validierung durch ScratT



2020 - 0502

6.1 Validierung durch ScraT

Die Validierung beginnt schon während der automatisierten Preiserhebung. ScraT ist in der Lage, bestimmte Charakteristika eines Produkts zu prüfen, bevor der dazugehörige Preis erhoben wird. Solche Charakteristika können allgemeiner Art sein, wie Artikelnummern der zu erhebenden Produkte oder produktspezifische Merkmale, beispielsweise die Verbindungsstrecken beim Personentransport. Dieses Vorgehen erhöht die Präzision der automatisierten Preiserhebung beträchtlich und reduziert die ungeeigneten Daten in der Ergebnisdatenbank. Dieser Validierungsprozessschritt wird programmtechnisch unterstützt. ↘ Grafik 7

Erklärung: Wenn die extrahierte Artikelnummer der Artikelnummer aus den Eingangsdaten entspricht, dann wird der Ablauf fortgesetzt.

6.2 Technische Validierung

Nach Abschluss der Preiserhebung erzeugt ScraT automatisch eine Datei mit einer Zusammenfassung des gerade abgeschlossenen Erhebungslaufs. In dieser werden Kennzahlen (zum Beispiel die Laufzeit der Erhebung, die Anzahl der erhobenen Preise, erkannte Fehler oder Warnungen) zusammengestellt. Darüber hinaus kann ein Video mit der Erhebung aufgezeichnet werden. Diese Informationen ermöglichen es den für den Ablauf der Erhebung verantwortlichen Mitarbeiterinnen und Mitarbeitern, die technische Wartung durchzuführen (siehe Kapitel 4). Bei Bedarf können sie eine Wiederholung der Erhebung starten, um entstandene Lücken zu schließen.

6.3 Fachliche Datenvalidierung

Nach einem aus technischer Sicht erfolgreichen Erhebungslauf kann die Datenvalidierung durch Produktexpertinnen und Produktexperten beginnen. Vorübergehend nicht verfügbare Produkte sind im stationären und im Internethandel normal. Um im Nachgang der Erhebung beurteilen zu können, ob ein Produkt nur vorübergehend nicht verfügbar war oder dauerhaft aus dem Sortiment genommen wurde, werden Informationen zur Verfügbarkeit des Produkts in die Ergebnisdatenbank abgelegt. Sollte ein Produkt mehr als zwei Monate nicht verfügbar sein, wird das Produkt ersetzt. Auch bei fehlerhaften vorgegebenen Positionen der zu erhebenden Informationen auf den Produktseiten oder gänzlich fehlerhaften URLs müssen die Produktexpertinnen und -experten die Datenbanken entsprechend pflegen und für die nächsten Erhebungen vorbereiten. Die erhobenen Preisdaten werden anschließend auf Ausreißer und nicht plausible Werte überprüft, beispielsweise mithilfe der statistischen Auswertungssoftware SAS.

6.4 Weiterverwendung von Web-Scraping-Preisindizes

Die monatliche Datenaufbereitung und Datenvalidierung der Verbraucherpreisstatistik insgesamt wird einheitlich in einem gemeinsamen Berechnungsprogramm der Statistischen Ämter des Bundes und der Länder durchgeführt. Dieses sogenannte Verbundprogramm[16]

16 Für zusätzliche Erläuterungen siehe Burg/Seeger (2009).

ist auf eine traditionelle Preiserhebung ausgerichtet und akzeptiert daher für die monatliche Indexberechnung nur einen einzelnen Preis oder eine Messzahl für das jeweilige Produkt. Messzahlen müssen außerhalb des Erhebungsprogramms berechnet und anschließend im Verbundprogramm erfasst werden. Diese Messzahlen können Lageparameter aus den erhobenen Daten sein, welche beispielsweise mit SAS berechnet werden.

7

Fazit und Ausblick

Vor der Einführung von ScraT wurde jeder Automatisierungsprozess individuell programmiert und musste auf isolierten Geräten mit separater Internetverbindung ablaufen. Jede automatisierte Erhebung wurde durch ein eigenes, individuelles Programm gesteuert, welches alle Erhebungsschritte durchlaufen lässt und die erhobenen Daten bearbeitet. So mussten für jeden Onlinehändler oder jede Suchmaschine eigene Programme entwickelt und gewartet werden. Das führte zu einer hohen Abhängigkeit von IT-Kenntnissen bei den handelnden Beschäftigten und zu einem hohen Wartungsaufwand. Mit der Entwicklung der Anwendung ScraT wird das automatisierte Auslesen von Daten im Internet und gleichzeitig die Verwendung von Web Scraping weiter optimiert. Die vorhandenen Web-Scraping-Programme werden standardisiert und breit nutzbar gemacht.

Die bisher überwiegend einmal monatlich für jedes Produkt durchgeführten Preiserhebungen im Internet werden zukünftig automatisiert und in vielen Fällen in einer sehr viel höheren Frequenz durchgeführt. Die bestehende Stichprobe der etwa 10 000 Produkte im Internethandel muss dabei nicht ausgeweitet werden, sondern entsprechend der Volatilität der Preise beziehungsweise der Preissetzungsstrategien der Onlinehändler werden mehr Preise je Produkt erfasst. Plausibilisierungsschritte, die bislang im Zuge der Preiserhebung erledigt wurden, müssen nachgelagert und sehr zeitnah im Erhebungsprozess für einen deutlich größeren Datensatz erfolgen. Insgesamt steigen vor allem durch die fachliche Datenvalidierung die qualitativen Anforderungen an die Mitarbeiterinnen und Mitarbeiter.

Die Fertigstellung der Entwicklung und der Beginn der Einführung des generischen Programms für Web Scraping, ScraT, stellen zusammen einen wichtigen Baustein in der Digitalisierung der amtlichen Statistik dar. ScraT liefert zudem einen wichtigen Beitrag zur Effizienz- und Qualitätssteigerung im Statistischen Bundesamt. Das Ziel ist, bis Ende des Jahres 2021 die vormals manuellen Preiserhebungen im Internet weitgehend zu automatisieren. Zunächst werden dabei die bisherigen manuellen Erhebungen imitiert und bei volatilen Preisen die Erhebungsfrequenz angepasst. Die Weitergabe des Programms an die Statistischen Ämter der Länder ist angestrebt. Langfristig ist es möglich, das generische Programm auch für die Datenerhebung anderer Statistiken zu erweitern und zu verwenden. Außerdem werden die Vor- und Nachteile von „bulk scraping“ (Abgriff von Massendaten) in einer zeitnahen Studie evaluiert.

ScraT unterstützt auch ein strategisches Ziel des Statistischen Bundesamtes, indem es zur Digitalisierung von Statistiken beziehungsweise Prozessen beiträgt. Erhobene Daten liegen standardisiert vor, erste Plausibilisierungen erfolgen bereits während der Erhebung und automatisierte Auswertungen in SAS oder der statistischen Programmiersprache R schließen sich an. Damit erhöht sich die Produktivität in der Statistikproduktion, die Datengrundlage wird erweitert und die Datenqualität verbessert. Zusätzlich fördert es die Forschung im Bereich der dynamischen Preissetzung im Internethandel, welche auf großes öffentliches Interesses stößt (Hansen, 2020a; Hansen, 2020b). Im digitalen Zeitalter sichert das Statistische Bundesamt somit das Vertrauen in seine Ergebnisse und steigert die Relevanz seiner Statistiken.

LITERATURVERZEICHNIS

Blaudow, Christian. *Fortschritte und Herausforderungen beim Web Scraping – Automatisierung von Preiserhebungen im Internet*. In: Methoden – Verfahren – Entwicklungen. Ausgabe 1/2018, Seite 3 ff.

Blaudow, Christian/Burg, Florian. *Dynamische Preissetzung als Herausforderung für die Verbraucherpreisstatistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2018, Seite 11 ff.

Brunner, Karola. *Automatisierte Preiserhebung im Internet*. In: Wirtschaft und Statistik. Ausgabe 4/2014, Seite 258 ff.

Burg, Florian/Seeger, Daniel. *Das neue Verbundprogramm der Verbraucherpreisstatistik*. In: *Wirtschaft und Statistik*. Ausgabe 2/2009, Seite 169 ff.

Fentem, Katharina. *E-Commerce in Corona-Zeiten: Das sind die Trends*. 2020. [Zugriff am 10. September 2020]. Verfügbar unter: www.onlinehaendler-news.de.

Hansen, Malte. *Dynamische Preissetzung im Onlinehandel: zur langfristigen Anwendung von automatisierter Preiserhebung*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2020, Seite 14 ff. (2020a)

Hansen, Malte. *Dynamische Preissetzung im Onlinehandel: zu den Auswirkungen auf den Verbraucherpreisindex*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2020, Seite 91 ff. (2020b)

Plewa, Werner. *So erstellen Sie gute User Stories*. 2019. [Zugriff am 10. September 2020]. Verfügbar unter: www.business-wissen.de

Sandhop, Karsten. *Geschäftstypengewichtung im Verbraucherpreisindex*. In: Wirtschaft und Statistik. Ausgabe 3/2012, Seite 266 ff.

Schwaber, Ken/Sutherland, Jeff. *Der Scrum Guide™*. 2017. [Zugriff am 10. September 2020]. Verfügbar unter: www.scrumguides.org

Statistisches Bundesamt. *Onlinehandel gewinnt immer mehr an Bedeutung*. 2020. [Zugriff am 9. September 2020]. Verfügbar unter: www.destatis.de

Stewart, Simon/Burns, David. *WebDriver Level 2, W3C Working Draft 24 August 2020*. 2020. Verfügbar unter: www.w3.org/TR/webdriver/

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktionsleitung: Juliane Gude
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Oktober 2020
Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Artikelnummer: 1010200-20005-4, ISSN 1619-2907