Kühnemann, Heidi

Anwendungen des Web Scraping in der amtlichen Statistik

AStA Wirtschafts- und Sozialstatistisches Archiv

Provided in Cooperation with:
Springer Nature

Suggested Citation: Kühnemann, Heidi (2021) : Anwendungen des Web Scraping in der amtlichen Statistik, AStA Wirtschafts- und Sozialstatistisches Archiv, ISSN 1863-8163, Springer, Berlin, Heidelberg, Vol. 15, Iss. 1, pp. 5-25, https://doi.org/10.1007/s11943-021-00280-5

AStA Wirtsch Sozialstat Arch (2021) 15:5–25
https://doi.org/10.1007/s11943-021-00280-5

ORIGINALVERÖFFENTLICHUNG

Anwendungen des Web Scraping in der amtlichen Statistik

Heidi Kühnemann

Eingegangen: 20. Juli 2020 / Angenommen: 22. Februar 2021 / Online publiziert: 23. März 2021

Zusammenfassung Im World Wide Web (kurz „Web“) sind große Datenmengen verfügbar, die auch die amtliche Statistik für sich nutzbar machen kann. Die Extraktion dieser Daten durch Web Scraping bietet vielfältige Potenziale, beispielsweise die Kosten für die Datenerhebung reduzieren, Befragte entlasten, die Qualität amtlicher Daten verbessern oder stichprobenrelevante Einheiten in Befragungen identifizieren. Am Beispiel der Preis-, Tourismus-, Arbeitsmarkt- und Unternehmensstatistik wird in diesem Beitrag aufgezeigt, wie die amtliche Statistik in Deutschland bereits Web Scraping nutzt. Viele der hier aufgeführten Anwendungen befinden sich noch in einem frühen Entwicklungsstadium. In anderen nationalen Statistikämtern werden Daten aus dem Web zum Teil bereits in einem größeren Ausmaß für experimentelle Statistiken und im Produktivbetrieb genutzt. Dies ist unter anderem auf eine teils unzureichende rechtliche Grundlage von Web Scraping in der amtlichen Statistik in Deutschland, auf eine für die Methode nicht adäquate IT-Infrastruktur sowie auf einen Mangel an Mitarbeitenden mit den notwendigen Qualifikationen zurückzuführen.

Schlüsselwörter Web Scraping · Neue digitale Daten · Europäisches Statistisches System

JEL C800

H. Kühnemann (✉)
Hessisches Statistisches Landesamt und Statistisches Bundesamt, Wiesbaden, Deutschland
E-Mail: heidi.kuehnemann@statistik.hessen.de

Applications for web scraping in official statistics

Abstract Accessing the increasing amount of data available on the World Wide Web ("web" in short) by means of web scraping offers new possibilities for official statistics. Possible benefits of automatically extracting data from the web include a reduction of costs for data collection, a decreased burden for respondents, an improved quality of statistical products as well as a more targeted approach to identifying units of interest for surveys. This article uses official statistics about prices, tourism, employment and enterprises as an example to showcase how web scraping is used in official statistics in Germany. Many of these applications are still in an early stage of development. In comparison, some other national statistical institutes use data from the web more intensely for experimental statistics as well as in production. The major reasons for the hesitant usage of web scraping in official statistics in Germany are the absence of a broad legal basis for web scraping, an inadequacy of the IT-infrastructure as well as a lack of personnel with the necessary qualifications.

Keywords Web Scraping · Big Data · European Statistical System

JEL C800

1 Einleitung

Das Web als Datenquelle kann große Potenziale für die amtliche Statistik erschließen. Die Nutzung von Web Scraping in der amtlichen Statistik als Methode zur automatisierten Extraktion von Daten aus dem Web (Mitchell 2018, S. IXf) lässt erhoffen, dass der Erhebungsaufwand aufgrund des neuen Informationszugangs reduziert wird und Auskunftspflichtige entlastet werden (Blaudow und Ostermann 2020). Zusätzlich können dadurch neue Analysewege, Erkenntnisse und Inhalte gewonnen werden, etwa, wenn Textklassifikation von Unternehmenswebseiten zur Untersuchung von E-Commerce genutzt wird (Peters 2018b) oder wenn neue Erkenntnisse über die Struktur des Online-Stellenmarktes durch die Analyse von Online-Stellenanzeigen gewonnen werden (Rengers 2018a). Auch die Qualität von amtlicher Statistik hat das Potenzial, durch Web Scraping weiter verbessert zu werden. Internetdaten können in hoher Frequenz erhoben und automatisiert verarbeitet werden, sodass daraus entstehende Statistiken an Aktualität gewinnen (Hackl 2016). Außerdem können mit Web Scraping oftmals (annähernd) Grundgesamtheiten erhoben werden, beispielsweise die Webseiten aller Unternehmen mit Webpräsenz in Deutschland. Auch ein Abgleich von Daten aus dem Web mit administrativen und Befragungsdaten ist denkbar, um die Qualität der jeweiligen Datenquelle zu überprüfen und die korrekte Ausprägung eines interessierenden Merkmals zu schätzen (Peters 2018a). Web Scraping kann auch zur Unterstützung von Befragungen genutzt werden, beispielsweise um Befragte auszuwählen, die von besonderem Interesse sind. In außergewöhnlichen Umständen, beispielsweise während der Corona-Krise 2020/21, in der Befragungen insbesondere von Unternehmen nur bedingt möglich waren, können Daten aus dem

Internet einen Beitrag dazu leisten, der Nachfrage nach aktuellen Statistiken nachzukommen (siehe beispielsweise Kinne et al. 2020).

Dieser Beitrag stellt unterschiedliche bereits existierende oder in der Entwicklung befindliche Anwendungen von Web Scraping in der amtlichen Statistik vor. Im folgenden Abschnitt wird zunächst darauf eingegangen, was Web Scraping ist und welche Daten damit gewonnen werden können. Im Anschluss wird ein Überblick über bereits existierende Anwendungsfelder von Web Scraping in der amtlichen Statistik gegeben. Danach wird auf Herausforderungen für die Nutzung von Web Scraping in der amtlichen Statistik in Deutschland eingegangen. Zum Abschluss wird ein Fazit gezogen.

2 Was ist Web Scraping?

Die automatisierte Extraktion von Daten aus dem Web wird als Web Scraping bezeichnet (Condron et al. 2019). Dabei wird in der Regel nicht die Webseite so gescrapt, wie sie für menschliche NutzerInnen unmittelbar sichtbar ist, sondern der dahinterliegende HTML-Code, der das Aussehen und die Inhalte der Webseite bestimmt, wird ausgewertet. Eine Webseite (engl. „webpage“) ist ein Dokument im Web, das aus strukturiertem Text besteht (zumeist HTML-Code) und mit einer Webadresse bzw. dem Uniform Resource Locator (URL) erreichbar ist. Eine Webpräsenz (engl. „website“) besteht zumeist aus mehreren dieser Webseiten, die alle den gleichen URL-Stamm, Domain genannt, haben (z. B. amazon.com, spiegel.de). Die Homepage einer Webpräsenz ist die Webseite, auf die der Browser navigiert, wenn man die Domain eingibt.

Web Scraping-Programme simulieren das Verhalten menschlicher NutzerInnen, um bestimmte Informationen auf Webseiten zu erhalten (für eine Einleitung siehe beispielsweise Mitchell 2018). Die Interaktion mit den Webseiten verläuft dabei ganz ähnlich wie bei menschlichen NutzerInnen. Die menschliche Interaktion mit dem Web erfolgt im Allgemeinen über einen Browser, der mit dem Server der Webseite kommuniziert und den HTML-Code visuell ansprechend darstellt. Beim Web Scraping werden teilweise Browser so programmiert, dass sie bestimmte Webseiten besuchen und dort festgelegte Bestandteile der Webseite extrahieren. Auch das Ausfüllen von Formularen (z. B. Such- oder Login-Felder) kann automatisiert werden. Web Scraping ist allerdings auch ganz ohne Browser möglich. Das Web Scraping-Programm kommuniziert dann ohne den Umweg über einen Browser mit dem Webseiten-Server. Der HTML-Code der Webseite wird im Anschluss lokal ausgewertet.

Im deutschen Sprachgebrauch wird häufig Screen Scraping synonym zu Web Scraping verwendet (von Schönfeld 2018, S. 25). Der Begriff des Screen Scrapings impliziert jedoch, dass die Darstellung einer Webseite auf dem Bildschirm extrahiert wird und nicht der dahinterliegende Code oder die Datenbank. Außerdem umfasst Screen Scraping auch andere Anwendungen, die die Darstellung auf digitalen Bildschirmen aufzeichnen, beispielsweise schließt es auch die Extraktion von Daten ein, die außerhalb des Webs vorliegen. Web Scraping ist deswegen der präzisere Begriff, für die Anwendungen, die in diesem Artikel vorgestellt werden.

Ein weiterer Begriff, der häufig alternativ zu Web Scraping genutzt wird, ist Web Mining. Web Mining ist ein Sammelbegriff für den gesamten Prozess der automatischen Exploration und Extraktion von Information aus dem Web (Kosala und Blockeel 2000). Dieser Begriff ist einerseits weiter gefasst als Web Scraping, da darunter auch die Analyse der Nutzung und Struktur des Web verstanden wird, nicht nur die Extraktion von Informationen aus den Inhalten von Webseiten (letzteres wird als Web Content Mining bezeichnet). Andererseits ist Web Content Mining wiederum sehr stark auf die Aufbereitung von unstrukturierten Daten, wie beispielsweise Text oder Bilder, fokussiert. Web Scraping beschränkt sich jedoch nicht nur auf diese unstrukturierten Daten, sondern wird auch genutzt, um sehr spezifische Informationen auf Webseiten abzurufen, für die nur wenige weitere Verarbeitungsschritte notwendig sind. Ein Beispiel hierfür ist die Extraktion von Preisen aus Online-Shops. Insgesamt haben die Begriffe Web Mining, Web Scraping und Screen Scraping eine große Schnittmenge und können in einigen Fällen synonym verwendet werden. In diesem Beitrag wird der Begriff Web Scraping genutzt.

Web Crawling ist ein weiterer Begriff, der häufig synonym zu Web Scraping verwendet wird. Während Web Scraping die Extraktion einer einzelnen Webseite bezeichnet, ist das Ziel von Web Crawling das Finden von Verlinkungen auf einer Webseite. Diesen Verlinkungen wird dann gefolgt, um weitere Links zu finden (Mitchell 2018, Kap. 3). Web Crawling wird vor allem von Suchmaschinen angewendet, die das Web durchsuchen und indizieren. Auch in der amtlichen Statistik gibt es Anwendungen von Web Crawling, etwa wenn die Webpräsenz eines Unternehmens inklusive aller ihrer Webseiten („webpages“) gescrapt werden soll, um den extrahierten Text für die Ableitung latenter Eigenschaften zu nutzen. Web Crawling ist dann der Prozess, um die Webseiten zu identifizieren, die gescrapt werden sollen.

Des Weiteren kann spezifisches und generisches Web Scraping voneinander unterschieden werden (Stateva et al. 2018). Bei spezifischem Web Scraping ist die Struktur der Webseite im Voraus bekannt und nur bestimmte Informationen werden zielgerichtet extrahiert. Dies ist beispielweise der Fall beim Web Scraping in der Preisstatistik. Preise befinden sich im selben Online-Shop in der Regel immer an der gleichen Stelle im HTML-Code und können deswegen direkt extrahiert werden. Im Gegensatz dazu ist beim generischen Web Scraping der Inhalt und die Struktur der Webseite nicht bekannt. In diesem Fall wird der gesamte Inhalt der Webseite extrahiert. Ein Anwendungsfall ist hierbei z. B. die Ableitung latenter Eigenschaften von Unternehmenswebseiten. Es ist nicht bekannt, wie einzelne Unternehmen ihre Webpräsenz aufbauen. Deswegen wird der gesamte HTML-Code extrahiert.

Ein bedeutsamer Aspekt des Web Scraping ist die Datenextraktion. Der lokal vorliegende HTML-Code ist noch unstrukturiert und stellt an sich noch keinen Erkenntnisgewinn dar. Beim spezifischen Web Scraping werden die Daten von Interesse meist anhand ihrer Position im HTML-Code extrahiert. HTML ist mit sogenannten „tags“ (dt. Auszeichnungsmarkierungen) strukturiert, die den Bestandteilen des Codes Funktionen und Eigenschaften zuweisen. Mit Hilfe dieser „tags“ können Teile des Codes sehr genau ausgewählt werden.

Beim generischen Web Scraping ist häufig der gesamte Text auf einer Webseite von Interesse. Dieser Text wird dann mit Methoden des Text Mining in eine Form gebracht, die für die statistische Analyse genutzt werden kann. Text Mining ist hierbei

ein Sammelbegriff verschiedener Algorithmus-basierter und statistischer Verfahren für die Analyse von Text-Daten (für eine Einführung siehe Gentzkow et al. 2019).

Manche Webseitenbetreiber bieten direkt einen automatisierten Zugriff auf die Datenbanken an, auf denen die Webpräsenz basiert. Dies geschieht in Form einer sogenannten Web-API (für Englisch „application programming interface"; Mitchell 2018, Kap. 12). Die Verwendung einer API ist dem regulären Web Scraping im Allgemeinen vorzuziehen, unter anderem weil eine API die interessierenden Daten in strukturierter Form übermittelt. Allerdings ist nur ein sehr geringer Teil der Informationen im Web über eine API verfügbar. Wenn im Folgenden Anwendungen des Web Scraping in der amtlichen Statistik vorgestellt werden, kann das auch Anwendungen einschließen, die zur Informationsbeschaffung eine API nutzen, da damit die gleiche Datenquelle erschlossen wird.

3 Anwendungsgebiete von Web Scraping

Im Folgenden werden aktuelle Anwendungsgebiete von Web Scraping in der amtlichen Statistik in Deutschland nach Themen- bzw. Fachbereichen aufgeführt. Die Reihenfolge, in der die Anwendungsbeispiele aufgeführt werden, entspricht dabei dem Grad der Implementierung in die Produktion. Zuerst werden im Folgenden Anwendungen aufgeführt, die bereits in der Statistikproduktion genutzt werden.

3.1 Preisstatistik

Web Scraping als Methode zur Erhebung von Preisen für den Verbraucherpreisindex ist bereits gut etabliert. Aufgrund der großen Bedeutung des Internethandels in Deutschland werden ungefähr 10.000 Produkte monatlich online – allerdings zum Teil noch manuell – für die Verbraucherpreisstatistik erhoben. Die Preise einiger Produktgruppen werden bereits vollständig automatisiert erfasst. Dies trifft beispielsweise auf Mietwagen sowie Fernbus- und Bahnreisen zu. Bis Ende 2021 ist es das Ziel, alle Produkte in der Online-Preiserhebung mittels Web Scraping automatisiert zu erheben. Aus diesem Grund wurde das generische Programm „ScraT" (Scraping-Tool) erstellt, das Web Scraping von Preisen auch für MitarbeiterInnen ohne tiefgreifende IT-Kenntnisse möglich macht (Blaudow und Ostermann 2020). Die Preiserfassung geschieht dadurch im Vergleich zur manuellen Erhebung schneller und kosteneffizienter. Auch die Häufigkeit der Preiserhebung kann nach Bedarf stark erhöht werden. Dies verbessert die Qualität der Verbraucherpreisstatistik.

Neben der Erhebung von Preisen für den Verbraucherpreisindex wird Web Scraping in der Preisstatistik auch zur Untersuchung von dynamischer Preissetzung im Online-Handel genutzt. Dynamische Preissetzung im Online-Handel stellt eine besondere Herausforderung für die Preisstatistik dar, denn die traditionelle monatliche Erhebung eines Preises je Produkt ist damit nicht mehr repräsentativ. Web Scraping wird eingesetzt, um die Häufigkeit von Preisänderungen und die Varianz von Preisen zu untersuchen. Es konnte gezeigt werden, dass einige Online-Händler stärker Gebrauch von dynamischer Preissetzung machen als andere und auch die Variation der Preise sich zwischen Händlern unterscheidet. Die Häufigkeit der Preiserhebung

sollte demnach vom Händler abhängig gemacht werden (Blaudow und Burg 2018). Auch das Ausmaß von stündlichen, täglichen und saisonalen Preisänderungen konnte mit Web Scraping untersucht werden (Hansen 2020a, b). Daraus leitete Hansen (2020b) Handlungsempfehlungen für den Zeitpunkt von automatisierten Preisabfragen ab, um Zeitpunkte starker Preisvolatilität zu vermeiden.

In der amtlichen Statistik bislang noch nicht untersucht wurde das Phänomen der personalisierten Preissetzung, bei der Preise an die vermutete Zahlungsbereitschaft der potenziellen KäuferInnen angepasst werden (Zander-Hayat et al. 2016). Merkmale für die Prognose der Zahlungsbereitschaft sind beispielsweise das genutzte Endgerät, zuvor besuchte Webseiten oder die Verweildauer auf der jeweiligen Webseite. Dies stellt die amtliche Statistik vor neue Herausforderungen, da nicht nur der Zeitpunkt und die Häufigkeit der Preiserhebung verändert werden müssen, sondern es müssten auch typische KäuferInnenprofile simuliert werden, sodass die tatsächlich bezahlten Preise erhoben werden können. Bislang geschieht dies noch nicht. In einer Studie aus dem Jahr 2018 nutzten jedoch nur sehr wenige der untersuchten Online-Shops personalisierte Preissetzung und dies auch nur in einem geringen Umfang (Dautzenberg et al. 2018).

Die Verwendung von Web Scraping ist eine sinnvolle Methode, um die Preiserhebung im Internet zu automatisieren. Zudem ist es eine effiziente Methode, die Preiserhebung im Internet auszudehnen, da mittelfristig weniger Personalressourcen eingesetzt werden müssen als bei der traditionellen Preiserhebung. Einigen neueren Entwicklungen, wie der dynamischen Preissetzung, kann durch häufige Preiserhebungen durch Web Scraping begegnet werden. Personalisierte Preissetzung wird bislang jedoch noch nicht untersucht, was die Repräsentativität der erhobenen Preise – vermutlich nur in geringem Maße – einschränkt.

3.2 Insolvenzbekanntmachungen

Eine bereits etablierte Nutzung von Web Scraping in der amtlichen Statistik ist auch das Scrapen von Insolvenzbekanntmachungen auf der Webpräsenz „insolvenzbekanntmachungen.de“, betrieben durch das Justizministerium des Landes Nordrhein-Westfalen (für das gesamte Bundesgebiet). Dort sind für die jeweils aktuellen letzten zwei Wochen alle Bekanntmachungen zu Insolvenzverfahren verfügbar und können automatisch extrahiert werden. Diese Insolvenzbekanntmachungen werden durch das Statistische Bundesamt (Destatis) einmal pro Woche gescrapt und mittels Text Mining nach interessierenden Stichworten sowie nach Aktenzeichen bzw. Gericht-Identifikatoren durchsucht.

Diese Anwendung des Web Scraping dient zur Qualitätssicherung der Insolvenzstatistik, da so sichergestellt werden kann, dass alle Insolvenzverfahren in der Statistik erfasst wurden. Zusätzlich dienen die so gewonnenen Daten als Schätzgrundlage für die vierteljährliche Unternehmensdemografie. Während der Corona-Krise 2020/21 konnten diese Daten genutzt werden, um Statistiken zu eröffneten Insolvenzverfahren mit gesteigerter Aktualität anzubieten.[1]

[1] Siehe hierzu beispielsweise die Pressemitteilung des Statistischen Bundesamtes vom 11. Mai 2020: https://www.destatis.de/DE/Presse/Pressemitteilungen/2020/05/PD20_163_52411.html.

Das Verfahren ist jedoch zurzeit davon abhängig, dass die Webpräsenz „insolvenzbekanntmachungen.de“ unverändert bleibt. Wenn sich im HTML-Code der Seite etwas Grundlegendes verändert, funktioniert das Web Scraping-Programm nicht mehr und muss an die Änderungen angepasst werden. Das ist besonders deswegen problematisch, weil eine uneingeschränkte Suche in den Insolvenzbekanntmachungen nur zwei Wochen nach Veröffentlichung möglich ist. Anpassungen am Web Scraping-Programm müssten dann unter hohem Zeitdruck erfolgen. Dieses Problem könnte dadurch gelöst werden, dass das Justizministerium NRW der amtlichen Statistik eine API zur Verfügung stellt oder eine regelmäßige Datenlieferung an die amtliche Statistik in einer anderen Form erfolgt.

3.3 Tourismusstatistik

Web Scraping wird im Hessischen Statistischen Landesamt (HSL) zur Anreicherung und Qualitätskontrolle der monatlichen Tourismusstatistik verwendet. Die hessische Beherbergungsstatistik erfasst monatlich etwa 3500 Betriebe mit mindestens 10 angebotenen Betten. Viele dieser Betriebe sind nicht nur im Besitz einer unternehmenseigenen Webpräsenz, sondern zusätzlich auf einem kommerziellen Onlineportal, beispielsweise Booking.com, HRS-Holiday oder Hotel.com präsent, auf dem sie ihre Dienstleistungen anbieten. Dort sind Informationen über die Unterkunftsanbieter mit vielen Attributen abrufbar und aufgrund der innerhalb des Onlineportal gleichbleibenden Struktur im Aufbau der zugehörigen Webseiten ohne aufwendige Textverarbeitungsschritte extrahierbar. Durch die anschließende Verknüpfung mit dem Berichtskreis der Beherbergungsstatistik kann ein Abgleich erfolgen und die Vollständigkeit des Berichtskreises überprüft werden. Die Angaben über die angebotene Zimmeranzahl ist in vielen tourismusbezogenen Onlineportalen erhältlich und dient zur Bestimmung des Bettenangebotes. Diese Größe ist in der Beherbergungsstatistik eines der Erhebungsmerkmale und ist außerdem entscheidend für die Auskunftspflicht des Betriebes. Daneben sind Hilfsmerkmale wie E-Mail-Adresse oder Telefonnummer ebenfalls häufig in den Onlineportalen verfügbar. Somit kann Web Scraping hier als bedeutende Unterstützung für Pflege und Aktualisierung der für die Statistik zu erfassenden Betriebe eingesetzt werden. Da es möglich ist, die Angebote auf tourismusbezogenen Onlineplattformen nach verschiedenen Regionen zu durchsuchen, sind diese auch für die bundesweite Tourismusstatistik eine vielversprechende Datenquelle.

Im Rahmen einer Machbarkeitsstudie ist es dem HSL gelungen, durch das automatisierte Extrahieren der Daten eines kommerziellen Online-Buchungsportals für Unterkünfte, 2438 hessische Beherbergungsbetriebe inklusive weiterer Informationen, wie die Zimmer- und Bettenanzahl, zu erheben (Peters 2018a). In diesem Rahmen konnten auch hessische Unterkunftsanbieter identifiziert werden, die aufgrund der angegebenen Zimmeranzahl im Grenzbereich der Auskunftspflicht lagen. Der Berichtskreis der Beherbergungsstatistik konnte somit um diese Betriebe ergänzt werden. Auch bei der Plausibilisierung des Umsatzes von Beherbergungsbetrieben könnten die automatisiert erhobenen Daten äußerst nützlich sein, da die Anzahl der angebotenen Schlafgelegenheiten im Allgemeinen positiv mit dem Unternehmensumsatz korreliert.

3.4 Online-Stellenanzeigen

Der Einsatz von Web Scraping für die Arbeitsmarktstatistik ist bereits seit mehreren Jahren Gegenstand der Forschung bei Destatis. Im Rahmen der Projekte ESSnet Big Data 2016–2018 und 2018–2020[2] wird die Möglichkeit untersucht, wie Online-Stellenanzeigen analysiert und daraus neue Indikatoren zum Stellenmarkt und zur Arbeitsnachfrage gewonnen werden können (Rengers 2018a, b).

Diese Stellenanzeigen werden vorrangig aus Online-Jobportalen bezogen. Eine Herausforderung ist dabei die große Zahl von Jobportalen in Deutschland. Für die Analyse werden in den Projekten zwei Datenquellen genutzt: einerseits Stellenanzeigen aus zwei Datenlieferungen der Bundesagentur für Arbeit (BA), die das größte deutsche Jobportal betreibt, und andererseits Daten des Europäischen Zentrums für Förderung der Berufsbildung (Cedefop[3]). Cedefop betreibt schon seit 2013 Web Scraping von Jobportalen in mehreren europäischen Ländern und verfügte bereits vor dem Projektstart über Erfahrung im Web Scraping, eine bestehende Infrastruktur sowie eine gute Ausstattung mit Personal und IT-Ressourcen. Es wurde daher entschieden, auf die bestehende Infrastruktur zurückzugreifen um die begrenzten Mittel der nationalen Statistikbehörden besser zu nutzen. Für die zweite Phase der Studie waren dort die Daten von 134 Jobportalen aus Deutschland verfügbar, welche über Marktbeobachtung ermittelt wurden.

Die Datenqualität ist in diesem Projekt eine besondere Herausforderung, unter Anderem da Informationen aus dem Fließtext von heterogenen Anzeigen gewonnen werden müssen. Stellenanzeigen werden häufig auf mehreren Portalen veröffentlicht und erscheinen mitunter sogar auf demselben Portal mehrmals. Gleichzeitig führt diese Redundanz aber auch zu einer besseren Abdeckung des Online-Stellenmarktes, da aus technischen Gründen, beispielsweise aufgrund von Veränderungen der Webseiten, die eine Anpassung der Web Scraping-Programme erfordern, als auch wegen nicht ausreichender Rechenkapazitäten, nicht alle Portale ununterbrochen erfasst werden können. Deswegen ist es notwendig, Dubletten zu identifizieren und zu bereinigen (Rengers 2018a). Bislang noch nicht untersucht wurde der Bias, der daraus entsteht, dass für BewerberInnen attraktive Stellenangebote nach kurzer Zeit wieder offline genommen werden, da bereits genügend Bewerbungen vorhanden sind. Die gescrapten Daten würden demnach überproportional viele unattraktive Stellenanzeigen beinhalten. Dieser Bias kann durch eine ausreichend hohe Frequenz an Scraping-Versuchen vermieden werden, da dann auch Stellenanzeigen mit einer sehr kurzen Verweildauer auf den Jobportalen erfasst werden. Zum jetzigen Zeitpunkt wurde jedoch noch nicht untersucht, ob die Häufigkeit des Scrapens durch Cedefop ausreicht, um diesen Bias zu minimieren.

Sowohl bei den Cedefop-Daten als auch bei den Daten der BA fehlt eine präzise Angabe zur Aktualität der Stellenanzeige, da das Veröffentlichungsdatum der Stellenanzeige entweder nicht erhoben werden kann oder nicht erhoben wurde. Auch bereits besetzte Stellen könnten im Datensatz enthalten sein. Einige für die Analy-

[2] Weitere Informationen unter https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data.

[3] Abkürzung steht für „Centre européen pour le développement de la formation professionnelle“.

se hochinteressante Merkmale der Stellenanzeigen, wie beispielsweise die Anzahl der Stellen pro Anzeige, sind insbesondere bei den Cedefop-Daten nicht enthalten (Rengers 2018a).

Es ist davon auszugehen, dass sich der Online-Stellenmarkt systematisch vom gesamten Stellenmarkt unterscheidet. Jedoch ist der Online-Stellenmarkt von großer Bedeutung in Deutschland, da laut der IAB-Stellenerhebung 2015 ungefähr 41 % aller Unternehmen Online-Jobportale als Mittel zur Rekrutierung neuer MitarbeiterInnen angeben. Eine Stellenanzeige auf der eigenen Webpräsenz ist sogar noch beliebter; 52 % aller Unternehmen gaben an, diesen Weg zu nutzen (Brenzel et al. 2016). Analysen aus der Schweiz bestätigen diese Einschätzung und legen zudem nahe, dass Anzeigen exklusiv über Printmedien kaum mehr eine Rolle spielen (Sacchi 2014). Eine Übersicht über die Struktur des Online-Stellenmarktes trägt demnach dazu bei, den Arbeitsmarkt in Deutschland besser zu verstehen. Die Daten der Bundesagentur für Arbeit konnten genutzt werden, um die Verteilung von Online-Stellenanzeigen und -Angeboten über Wirtschaftszweige, Berufe und Betriebsgröße zu untersuchen. Eine Herausforderung war hierbei, dass die Definitionen dieser Merkmale der Bundesagentur für Arbeit von den in der Statistik gebräuchlichen Definitionen abweichen und damit ein Vergleich zum gesamten Arbeitsmarkt für manche Merkmale nur eingeschränkt oder gar nicht möglich war. Der Cedefop-Datensatz orientiert sich dagegen an gängigen internationalen Klassifikationen.

Für die Fortführung der Untersuchung im Rahmen des ESSnet Big Data II wird der Online-Stellenmarkt im Längsschnitt untersucht. Für diese Fortführung werden Cedefop-Daten verwendet, für die im Vergleich zu der Studie von Rengers (2018a) wesentlich mehr Jobportale einbezogen wurden und bei der die Frequenz des Scrapings erhöht wurde, sodass der deutsche Online-Stellenmarkt wesentlich vollständiger abgebildet werden kann.

Im Rahmen der zweiten Projektphase werden zwei Arten von Indikatoren auf Basis von Online-Stellenanzeigen entwickelt. Einerseits sind dies monatliche Indizes zur Arbeitsnachfrage nach Region, Wirtschaftszweig und Qualifikation, vergleichbar dem Internet Vacancy Index des Australischen Arbeitsministeriums (Australian Government 2020). Die Daten von Cedefop finden dafür Verwendung. Da die genaue Gültigkeitsdauer einzelner Stellenanzeigen nicht bekannt ist, wird ein synthetischer Bestand auf Basis durchschnittlicher Gültigkeit errechnet. Außerdem werden Indizes errechnet, die die relative Veränderung der Arbeitsnachfrage gegenüber dem Startwert abbilden, weil trotz Deduplizierung höchstwahrscheinlich Anzeigen mehrfach enthalten sind.

Andererseits wird ein Indikator zur Arbeitsmarktkonzentration auf Nachfrageseite erstellt. Dieser zeigt, wie viele Unternehmen pro Region Stellenangebote für einen bestimmten Beruf anbieten. Wenn nur wenige potenzielle Arbeitgeber in einer Region Stellen in bestimmten Berufen anbieten, so haben diese Arbeitgeber Marktmacht und können niedrigere Löhne anbieten. Online-Stellenanzeigen können somit zur Identifikation von Monopsonen oder Oligopsonen im Arbeitsmarkt genutzt werden (Azar et al. 2018).

3.5 Extraktion von Unternehmenseigenschaften anhand der Unternehmens-Webpräsenz

Ein großer Teil von Unternehmen in Deutschland hat eine eigene Webpräsenz. Diese Webpräsenz kann als Informationsquelle über Unternehmen genutzt werden, da von ihr zahlreiche Eigenschaften des Unternehmens abgeleitet werden können. Viele dieser Eigenschaften sind latent, da auf den Unternehmenswebseiten nur Merkmale direkt beobachtet werden können, die auf solche Eigenschaften hinweisen, jedoch in der Regel nicht die Eigenschaften selbst. Dazu zählen beispielsweise innovative Aktivitäten, gemeinnütziges Engagement oder die Zugehörigkeit zu einer bestimmten Branche. Ist der Zusammenhang zwischen dem Auftreten von solchen Merkmalen auf der Unternehmens-Webpräsenz und der latenten Eigenschaft bekannt, so kann für jede unbekannte, automatisiert extrahierte Unternehmens-Webpräsenz nach Zuordnung bestimmt werden, ob ein Unternehmen über diese latente Eigenschaft verfügt (siehe beispielsweise Kühnemann et al. 2020). Web Scraping kann demnach einen Beitrag zur Anreicherung, Aktualisierung und Überprüfung verfügbarer Unternehmensdaten leisten.

Die Voraussetzung für die Anreicherung von Unternehmensdaten mit neuen Internetinhalten ist die Kenntnis über die URL der jeweiligen unternehmenseigenen Webpräsenz. Diese liegt jedoch nicht im Unternehmensregister vor. Beim Unternehmensregister handelt es sich um eine statistische Datenbank, die auf die vollständige Aufführung aller wirtschaftlich aktiven Unternehmen im deutschen Inland abzielt. Nach dem Vorbild eines vom Italienischen Statistischen Amt (ISTAT) publizierten Verfahrens (Barcaroli et al. 2016), wurde im HSL ein Algorithmus entwickelt, welcher über die Adressdaten einer Unternehmensdatenbank mit Hilfe einer Suchmaschine, wie beispielsweise Google, URLs sucht und den passenden Unternehmen zuordnet. Mit diesem Verfahren wurden bereits etwa 1200 Webpräsenzen von Unternehmen der hessischen Erhebung in Informations- und Kommunikationstechnologie IKTU 2017 gefunden und zugeordnet. Etwa 90 % der Zuordnungen stellten sich nach manueller Überprüfung einer Stichprobe von 100 Unternehmen als korrekt heraus (Peters 2018b).

Ausgehend von bekannten Unternehmens-Domains können diese Webpräsenzen gescrapt werden und als Grundlage für die Extraktion latenter Eigenschaften genutzt werden. Ein Beispiel für eine bereits existierende Anwendung in diesem Bereich ist eine Studie zur Messung der Internetökonomie in den Niederlanden (Oostrom et al. 2016). Dort wurden Daten von Unternehmens-Webpräsenzen als Grundlage verwendet, um Unternehmen in Kategorien der Internetnutzung zu klassifizieren. Diese Kategorien umfassten beispielsweise das Vorhandensein eines Online-Shops oder das Angebot von Online-Services. Es konnte gezeigt werden, dass 4,4 % aller Arbeitsstellen und 7,7 % des Gesamtumsatzes der Niederlande in der Internetökonomie zu verorten sind.

Das HSL arbeitet seit 2019 in eine ähnliche Richtung, da es untersucht, wie der Betrieb eines Online-Shops auf deutschen Unternehmens-Webpräsenzen automatisiert nachgewiesen werden kann. In einem ersten Versuch mit 8422 recherchierten Webpräsenzen von Unternehmen aus verschiedenen Registern konnten 86 % aller Unternehmen mit Online-Shop mit dem Klassifikationsalgorithmus korrekt identi-

fiziert werden. Etwa 10 % wurden dabei fälschlich als Online-Shop betreibendes Unternehmen klassifiziert.

Wenn URLs von Unternehmen verfügbar sind, können je nach Bedarf auch ganz neue Merkmale untersucht werden. Im Zentrum für Europäische Wirtschaftsforschung beispielsweise wurde schon kurz nach der Einführung weitreichender Kontaktbeschränkungen aufgrund der Corona-Pandemie in Deutschland analysiert, wie stark deutsche Unternehmen von der Krise betroffen waren. Dafür wurde untersucht, welche Unternehmen auf ihrer Webpräsenz die Corona-Krise erwähnen und in welchem Kontext dies geschah, beispielsweise ob Unternehmen schließen mussten oder nur die Öffnungszeiten anpassten (Kinne et al. 2020). Die Webseiten wurden mehrmals pro Woche gescrapt, um die zeitliche Entwicklung nachvollziehen zu können. Auch regionale Unterschiede wurden untersucht. Dies zeigt, dass der Zugang zu Unternehmens-URLs das Potenzial bietet, schnell auf neuartige Informationsbedarfe eingehen zu können.

3.6 Verlagerung wirtschaftlicher Aktivitäten ins Ausland

Eine ganz neue, momentan nur konzeptionell vorliegende Anwendung von Web Scraping in der Unternehmensstatistik betrifft die Identifikation von Unternehmen, die Geschäftsbereiche in das Ausland verlagern. Diese Unternehmen sind im Fokus einer ab den Berichtsjahren 2021–2023 einzuführenden Statistik zu globalen Wertschöpfungsketten. Die amtliche Statistik in Deutschland steht dabei vor der Herausforderung, dass mit knapp 2 % nur ein sehr kleiner Anteil der Unternehmen in der Grundgesamtheit tatsächlich wirtschaftliche Tätigkeiten in das Ausland verlagert (Kaus 2019a). Im Rahmen einer zufälligen Stichprobe ist der Großteil des Fragebogens für die Mehrheit der befragten Unternehmen nicht relevant. Dies wirkt sich vermutlich negativ auf die Akzeptanz und die Rücklaufquote der Erhebung aus, die beispielsweise bei der Befragung im Jahr 2016 sehr gering war (Kaus 2019b). Wenn Web Scraping genutzt werden könnte, um verlagernde Unternehmen bereits vor der Ziehung der Stichprobe zu identifizieren, könnten diese wesentlich zielgerichteter in die Stichprobe einbezogen werden. Damit würden sowohl die Auskunftslast der befragten Unternehmen als auch die Erhebungskosten verringert werden. Eine qualitative Analyse verschiedener online verfügbarer Datenquellen zeigte bereits, dass die Identifikation von Unternehmen mit Auslandsverlagerungen grundsätzlich möglich ist. Für die Identifikation von verlagernden Unternehmen müssten Unternehmensnamen aus dem statistischen Unternehmensregister (URS) genutzt und die Ergebnisse gescrapt werden.

3.7 Zusammenfassung der bisherigen Erfahrungen

Diese sechs Beispiele illustrieren das große Potenzial von Web Scraping in der amtlichen Statistik. Die Beispiele zeigen auf, dass Web Scraping in der amtlichen Statistik in mindestens drei verschiedenen Bereichen eingesetzt werden könnte:

- Datenerhebung,
- Qualitätssicherung und Plausibilisierung
- sowie die Ermittlung der Grundgesamtheit bzw. stichprobenrelevanter Einheiten.

Tab. 1 gibt eine Übersicht, welche Funktionen Web Scraping in den Beispielen aus diesem Beitrag erfüllt.

Des Weiteren ist interessant, welche unterschiedlichen Arten von Internet-Quellen gescrapt werden. Die Art der Webseite, die als Datenquelle genutzt wird, beeinflusst stark, welche Herangehensweise für das Web Scraping und die Datenaufbereitung genutzt wird. Hierbei ist vor allem der Strukturierungsgrad der Internetquelle von Bedeutung. Bei Webseiten mit einem gleichen Aufbau des HTML-Codes, auf denen die interessierende Information in einem gleichbleibenden Format immer auf derselben Position zu finden ist, kann von einem hohen Strukturierungsgrad gesprochen werden. Bei Webseiten, die alle einen unterschiedlichen Aufbau des HTML-Codes aufweisen und auf denen die interessierende Information unstandardisiert, beispielsweise als freier Text in diversen Formaten, vorliegt, kann von einem niedrigen Strukturierungsgrad gesprochen werden. Tab. 2 zeigt beispielhaft einige Webseiten-Arten sowie deren (ungefähren) Strukturierungsgrad auf und gibt an, wie diese für die amtliche Statistik genutzt werden. Als teilstrukturiert eingeschätzte Internetquellen haben häufig einen gleichbleibenden Aufbau des HTML-Codes, die interessierende Information liegt jedoch noch nicht standardisiert vor. Ein Beispiel ist die Adresse eines Beherbergungsbetriebes, die zwar immer auf der gleichen Position in einem Online-Buchungsportal steht, jedoch in unterschiedlichen Schreibweisen oder mit Rechtschreibfehlern vorliegt.

Im Moment werden in der deutschen amtlichen Statistik vorrangig Webseiten gescrapt, die bereits einen hohen Strukturierungsgrad besitzen. Dies sind einerseits

Tab. 1 Übersicht über Funktionen von Web Scraping in der amtlichen Statistik mit Beispielen aus Deutschland

Funktion	Beispiele
Datenerhebung	Preiserhebung im Online-Handel
	Beherbergungsstatistik
	Extraktion von Unternehmenseigenschaften
	Online-Stellenanzeigen
Qualitätssicherung und Plausibilisierung vorhandener Daten aus Befragungen oder administrativen Quellen	Insolvenzbekanntmachungen
	Beherbergungsstatistik
	Extraktion von Unternehmenseigenschaften
Ermittlung der Grundgesamtheit bzw. stichprobenrelevanter Einheiten	Beherbergungsstatistik
	Statistik zu globalen Wertschöpfungsketten/Verlagerungen ins Ausland

Tab. 2 Arten von Webseiten mit Beispielen für die Nutzung als Datenquelle in der amtlichen Statistik

Webseiten-Art	Strukturierungsgrad der Webseite	Beispielhafte Nutzung für die amtliche Statistik
Online-Shops	Strukturiert	Extraktion von Preisen im Onlinehandel
Suchmaschinen-ergebnisse	Teilstrukturiert	Finden von Unternehmens-URLs
Internetportale	Teilstrukturiert	Beherbergungsstatistik
		Insolvenzbekanntmachungen
		Online-Stellenanzeigen
Soziale Medien	Teilstrukturiert	Messung sozialer Spannungen auf Basis von Posts in sozialen Medien
Unternehmens-Webpräsenz	Unstrukturiert	Ableitung latenter Eigenschaften von Unternehmen, z. B. die Teilnahme am Online-Handel

Online-Shops, bei denen Produkt-Preise im Allgemeinen auf der gleichen Position im HTML-Code der jeweiligen Webseite gelistet werden. Die Preise liegen in einem oder einigen wenigen standardisierten Formaten (z. B. als Zahl mit maximal zwei Nachkommastellen und einem Euro-Zeichen) vor, sodass die Daten ohne oder nur mit minimalen weiteren Bearbeitungsschritten übernommen werden können. Auch Insolvenzbekanntmachungen im Internet haben einen gleichbleibenden Aufbau, in denen feststehende juristische Formulierungen genutzt werden, um die interessierende Information darzustellen. Die Beherbergungsstatistik profitiert ebenfalls davon, dass es nur relativ wenige Online-Buchungsportale gibt, innerhalb derer die Angebote für Unterkünfte gleichbleibend strukturiert sind. Diese gleichbleibende Struktur macht es möglich, ohne den Einsatz von maschinellen Lernverfahren oder aufwendiger Aufbereitung von Textdaten die interessierenden Daten zu extrahieren.

Die Nutzung von Quellen mit einem niedrigeren Strukturierungsgrad ist momentan in der deutschen amtlichen Statistik noch kaum etabliert. Die Ableitung von latenten Eigenschaften auf Basis der Unternehmens-Webpräsenz steht noch ganz am Anfang, insbesondere, weil Unternehmens-URLs im URS noch nicht verfügbar sind. Auch gescrapte Daten aus sozialen Medien werden momentan in der deutschen amtlichen Statistik noch nicht genutzt. Im europäischen statistischen System gibt es dafür durchaus Vorbilder. Das niederländische Amt für Statistik (CBS) bietet unter der Rubrik „experimentelle Daten“ ein Dashboard zu sozialen Spannungen und Emotionen im Zusammenhang mit Sicherheit und Gerechtigkeit an (Daas und Puts 2014). CBS erwirbt hierfür Posts aus zahlreichen sozialen Medien, vorrangig Twitter und Facebook, die bereits auf ihr Sentiment ausgewertet wurden. Sentiment bedeutet hierbei, ob Posts eine positive, negative oder neutrale Emotion transportieren. Ergebnisse werden innerhalb von 24 h online bereitgestellt[4]. Es konnte gezeigt werden, dass diese aggregierten Emotionen einen starken Zusammenhang mit dem Index für das Verbrauchervertrauen haben (Daas und Puts 2014).

Trotz all der aufgezeigten Vorteile befindet sich Web Scraping in Deutschland – mit Ausnahme der Preisstatistik – noch am Anfang. Im folgenden Abschnitt wird näher auf die Gründe dafür eingegangen.

[4] Zu finden unter: https://dashboards.cbs.nl/beta/experimenteel_SocialeSpanningen/ (auf niederländisch).

4 Herausforderungen

Die meisten der oben aufgeführten Beispiele für Anwendungen von Web Scraping in der amtlichen Statistik in Deutschland werden derzeit noch nicht in der regulären Produktion genutzt. Der Grund hierfür ist, dass zahlreiche Herausforderungen das Durchführen von Web Scraping in der amtlichen Statistik in Deutschland erschweren. Dies sind im Wesentlichen unzureichende oder nicht vorhandene einzelstatistische Regelungen, eine unzureichende IT-Infrastruktur sowie fehlende Qualifikationen der Mitarbeitenden für diese neue Methode.

4.1 Rechtliche Rahmenbedingungen

Öffentlich frei verfügbare Informationen, also beispielsweise Daten aus Web Scraping, dürfen durch die amtliche Statistik in Deutschland erhoben und für die Erstellung von Wirtschafts- und Umweltstatistiken mit Daten aus anderen Quellen verknüpft werden. Nach § 5 Abs. 5 Bundesstatistikgesetz (BStatG) dürfen die statistischen Ämter Daten aus „allgemein zugänglichen Quellen“ auch ohne Gesetz oder Rechtsverordnung erheben. Zur Pflege und Führung des Statistikregisters dürfen zudem gemäß § 13 Abs. 2 Satz 4 BStatG Angaben aus allgemein zugänglichen Quellen verwendet werden. Zudem erlaubt das BStatG die Zusammenführung von Daten aus Befragungen, Statistikregistern und sonstigen öffentlich verfügbaren Quellen (§13a BStatG) für die Wirtschafts- und Umweltstatistik. Diese Regelung hat insbesondere das Ziel, die Belastung von Auskunftspflichtigen zu reduzieren.

Eine Herausforderung beim Web Scraping für die amtliche Statistik stellen technische Barrieren auf Webseiten dar, die den Zweck haben, den Zugang für Web Scraping-Programme (sogenannte „Bots“) zu unterbinden. Diese technischen Barrieren dienen beispielsweise dazu, zu häufige Anfragen von Bots, die die Funktionsweise der Webseite gefährden könnten, zu verhindern. Selbstverständlich ist es nicht im Interesse der amtlichen Statistik, den normalen Betrieb von Webseiten durch Web Scraping zu stören. Um dies zu verhindern, wurden im ESSnet Big Data II Richtlinien zum ethischen Web Scraping erstellt, damit ein verantwortungsbewusster Umgang mit der Methode sichergestellt wird (Condron et al. 2019). Dies beinhaltet beispielsweise, die Belastung der WebseitenbetreiberInnen zu minimieren (etwa indem nicht zu viele Anfragen in kurzer Zeit an die gleiche Webseite gestellt werden), sich gegenüber der Webseite als Statistikamt zu identifizieren (über den sogenannten „user agent string“) sowie transparent über Methoden und Prozesse beim Web Scraping bzw. der Datenaufbereitung zu informieren[5].

Technische Barrieren, die Bots den Zugang zu Webseiteninhalte verwehren, erschweren der amtlichen Statistik auch unter Beachtung der Richtlinien zum ethischen Web Scraping das Erheben von Web-Daten. Die Preisstatistik ist hierbei eine Ausnahme, da sie bereits über eine einzelstatistische Rechtsgrundlage für den Einsatz von Web Scraping verfügt. Das Preisstatistikgesetz (PreisStatG) erlaubt seit dem 01.01.2020 ausdrücklich die Nutzung „automatisierter Abrufverfahren“ für allgemein zugängliche Preisinformationen und verpflichtet die „Halter der Daten [...] den

[5] Siehe beispielhaft die Umsetzung durch das HSL: https://statistik.hessen.de/ua/.

Abruf der Daten zu gewähren“ (§7b PreisStatG). Der Gesetzgeber erkennt hiermit an, dass die Nutzung von Web Scraping als neuer Erhebungsweg die Qualität der Preisstatistik sichern oder sogar verbessern kann (Deutscher Bundesrat 2019). Die Gesetzesänderung wird unter anderem damit begründet, dass der Handel im Internet stark zugenommen hat und sich im Zuge dessen auch die Preispolitik von Unternehmen verändert hat. Dies drückt sich vor allem in einer stärkeren Preisvolatilität aus. Das neue Preisstatistikgesetz befähigt die statistischen Ämter des Bundes und der Länder unter anderem, mit Web Scraping die Repräsentativität der erhobenen Preise zu sichern. WebseitenbetreiberInnen können im Zuge dessen die Statistikämter nicht daran hindern, Preisdaten automatisiert zu extrahieren. Das führt zu einer starken Erleichterung von Web Scraping für die Preisstatistik, da WebseitenbetreiberInnen nicht mehr mit technischen Mitteln die Scraping-Aktivitäten unterbinden können. Es existiert noch keine vergleichbare Regelung in anderen Statistikgesetzen.

4.2 IT-Infrastruktur

Die benötigte Hardware für Web Scraping ist abhängig von der Art der Anwendung. Für viele kleinere Projekte, die auf eine einmalige Datenerhebung abzielen, reicht ein einfacher Laptop[6] mit freiem Zugang zum Internet aus. Für die Implementierung von Web Scraping in die Statistikproduktion steigen die Anforderungen an die benötigte Hardware jedoch in Abhängigkeit vom Umfang, der Komplexität und den eingesetzten Methoden. Das Herunterladen von Webseiten und mehr noch die Aufbereitung großer Mengen unstrukturierter Daten, wie sie beispielsweise bei der Verarbeitung von Unternehmens-Webpräsenzen entstehen, stellen hohe Anforderungen an Arbeitsspeicher sowie die Zahl und Leistung der Prozessorkerne. Eine zukunftsfähige IT-Infrastruktur für Web Scraping würde auch leistungsfähige Grafikprozessoren einschließen, mit dem künstliche neuronale Netzwerke trainiert werden können. Künstliche neuronale Netzwerke kristallisierten sich als besonders geeignet für die Verarbeitung von Textdaten heraus (siehe beispielsweise Yang et al. 2016).

Anforderungen an die benötigte Software sind sowohl von dem Anwendungsziel als auch von den Erfahrungen der Mitarbeitenden abhängig. Stateva et al. (2018) stellten fest, dass, zumindest im Anwendungsgebiet Web Scraping von Unternehmensmerkmalen, die notwendige Web Scraping-Methodologie in zahlreichen Programmiersprachen bzw. mit der Unterstützung unterschiedlichster Software implementiert werden kann. Die Erstellung des Programms oder Skripts zur Durchführung des Web Scrapings kann beispielsweise in Java, Python oder R geschehen. Zusätzlich wird häufig eine Datenbank-Software für ein effizientes Management der gescrapten Daten benötigt. Hierfür sind – je nach Strukturierungsgrad der Datenquelle – sowohl relationale Datenbanken (beispielsweise MySQL) als auch nicht-relationale Datenbanken (beispielsweise MongoDB) geeignet. Software für Web Scraping ist oft Open-Source und häufig kostenfrei.

Die Wahl der genutzten Software kann einerseits davon abhängig gemacht werden, womit Mitarbeitende bereits Erfahrung haben, andererseits behindert eine Viel-

[6] Beispielhafte Konfiguration: Intel core i5, 2,4 GHz Taktfrequenz, 2 Kerne, 16 GB RAM.

zahl von unterschiedlicher Software die Möglichkeiten des Austauschs und die Nutzung von Synergieeffekten. Wenn unterschiedliche Software für die Entwicklung von Web Scraping-Anwendungen zur Verfügung steht, fördert das die Innovativität, da so auch völlig neue Methoden schnell umgesetzt werden und bereits bestehende Programmierkompetenzen der Mitarbeitenden genutzt werden können. Andererseits führt eine Nutzung zu vieler verschiedener Programmiersprachen dazu, dass gesammelte Erfahrungen nicht über eine bestimmte Anwendung hinaus nutzbar sind. Außerdem muss dann der Umgang mit Fehlermeldungen, die Erstellung von Logs und ähnliches für jedes Web Scraping-Programm individuell geregelt werden. Im Idealfall würden einige gängige IT-Tools und Programmiersprachen für Web Scraping identifiziert werden, mit der Web Scraping-Programme möglichst generalisiert entwickelt werden können, um die Chance für deren Wiederverwertbarkeit zu erhöhen.

Um Machbarkeitsuntersuchungen zur Nutzung von Web Scraping in der amtlichen Statistik in Deutschland und die Implementierung in die Produktion zu ermöglichen, wird ein Server mit freiem Zugang zum Internet und skalierbarer Rechen- und Arbeitsspeicherkapazität benötigt. Auch eine automatisierbare Schnittstelle dieses Web Scraping-Servers mit dem gesicherten Netz des jeweiligen Statistikamtes sollte vorhanden sein, um die Übertragung der erhobenen Daten zu gestatten. Die dafür erforderliche IT-Infrastruktur in den deutschen statistischen Ämtern existiert momentan noch nicht oder befindet sich noch im Aufbau.

4.3 Beschäftigtenqualifikation

Web Scraping ist keine Methode, die in gängigen Studiengängen der Wirtschafts- und Sozialwissenschaften gelehrt wird. Auch im Data Science-Studium ist Web Scraping in Deutschland kein zentraler Bestandteil im Curriculum. Web Scraping verlangt Beschäftigten gute Programmierfähigkeiten, gegenwärtig beispielsweise in R, Python oder Java, und Informatik-Kenntnisse, beispielsweise über die Parallelisierung von Prozessen oder Systemadministration, ab. Diese Fähigkeiten sind nur selten in Kombination mit einem ausreichend tiefen Verständnis für die Fachstatistiken, in denen die aus dem Web extrahierten Daten benötigt werden, anzutreffen. Projektgruppen für die Entwicklung von Web Scraping-Anwendungen, mit Mitgliedern aus der Fachstatistik und der IT (sowie optimalerweise ergänzt durch einen Data Scientist), sodass die notwendigen Kenntnisse nicht in einer Person vereint sein müssen, sind momentan in der amtlichen Statistik noch selten. Mitarbeitende, die Web Scraping in der amtlichen Statistik betreiben, haben sich die dafür nötigen Kenntnisse häufig autodidaktisch angeeignet. Fortbildungen in dem Feld sind zumeist Online-Kurse, die nicht auf die Besonderheiten in der amtlichen Statistik zugeschnitten sind. Einige Kurse des European Statistical Training Programme (ESTP) behandeln mittlerweile jedoch Aspekte des Web Scrapings, insbesondere geben Sie eine Übersicht über Anwendungsbeispiele und existierende IT-Tools zur Verarbeitung großer Datenmengen[7]. Das Angebot deckt jedoch nicht alle notwendigen Qualifikationen in ausreichender Tiefe ab.

[7] Siehe die Kurse „Introduction to Big Data in Official Statistics“ und „Big Data Tools for Data Scientists“ auf der ESTP-Webseite (https://ec.europa.eu/eurostat/cros/content/estp-training-offer_en).

5 Fazit und Ausblick

In diesem Beitrag wurden Anwendungsgebiete für Web Scraping in der amtlichen Statistik aufgezeigt. Die Methode bietet neue Potenziale und kann die Qualität amtlicher Statistik verbessern. Web Scraping senkt auch die Belastung für Befragte, da Informationen, die im Internet verfügbar sind, in Zukunft womöglich nicht mehr über einen Fragebogen erhoben werden müssen. Auch die Erhebungskosten können durch Web Scraping gesenkt werden, beispielsweise, weil Preise aus Online-Shops nicht manuell gesammelt werden müssen. Insgesamt kann erwartet werden, dass in Zukunft weitere Anwendungsgebiete des Web Scraping für die amtliche Statistik erschlossen werden. Trotz der vielfältigen Potenziale von Web Scraping in der amtlichen Statistik kann diese Methode im deutschen statistischen System bislang nur in einem Fall verbindlich eingesetzt werden. Die meisten Anwendungen befinden sich noch in einem frühen Entwicklungsstadium. Dies ist – mit Ausnahme der Preisstatistik – auf fehlende einzelstatistische Rechtsgrundlagen zurückzuführen. Außerdem ist die IT-Infrastruktur im deutschen statistischen System nicht auf die Anforderungen von Web Scraping ausgelegt. Zusätzlich gibt es einen Mangel an Beschäftigten mit den dazu notwendigen Kenntnissen und Fähigkeiten.

Auch Einschränkungen in der Qualität von Daten aus Web Scraping sind eine Herausforderung für die Nutzung dieser Daten in der amtlichen Statistik. An Web-Daten werden dabei in der amtlichen Statistik die gleichen Qualitätsanforderungen gestellt wie an Daten aus Befragungen oder administrativen Quellen. Qualitative Aspekte der gescrapten Daten müssen für jede Anwendung gesondert geprüft werden. Einschränkungen in der Qualität von Daten aus Web Scraping und anderen neuen digitalen Daten wurden beispielsweise im ESSnet Big Data II für die unterschiedlichen Anwendungen untersucht, sowie Richtlinien für den Umgang damit aufgestellt (Quaresma et al. 2020). Da nicht ausreichende Repräsentativität eine besondere Herausforderung für die Nutzung von neuen digitalen Datenquellen darstellt, wurde dieser Qualitätsaspekt ausführlich behandelt in Beręsewicz et al. (2018).

In einigen nationalen Statistikämtern im ESS wurden diese Herausforderungen bereits weitestgehend gemeistert. Als Beispiel ist hier CBS zu nennen, das bereits seit Jahren Daten aus dem Web für experimentelle Statistiken nutzt (vgl. z. B. Daas und Puts 2014; Oostrom et al. 2016). Allerdings ist diesen frühen Studien gemein, dass Daten, die mit Web Scraping gewonnen wurden, von global agierenden kommerziellen Unternehmen durch CBS erworben wurden. Für Daas und Puts (2014) wurde ein Datensatz mit Milliarden von Posts in Sozialen Medien gekauft. Die erworbenen Daten enthielten auch bereits das interessierende Merkmal, denn alle Posts wurden durch ein kommerzielles Unternehmen nach ihrem Sentiment klassifiziert. Es waren deswegen keine weiteren Textklassifikationsschritte notwendig. Auch Oostrom et al. (2016) nutzten Daten, die bereits zu einem hohen Grad für die Analyse aufbereitet waren. Dadurch beschränkte sich die Aufgabe der Autorinnen auf die Verknüpfung der Web-Daten mit amtlichen Datenquellen und auf traditionelle statistische Analysen – also auf Kernkompetenzen in der amtlichen Statistik.

Neben der eigenen Erhebung mittels Web Scraping ist der Erwerb von gescrapten und – zumindest zum Teil – aufbereiteten Daten eine Möglichkeit, um die Potenziale der im Web vorliegenden Datenmengen nutzen zu können. Zahlreiche global agie-

rende Unternehmen betreiben Web Scraping in großem Ausmaß und verfügen über potenziell interessante Daten für die amtliche Statistik (beispielsweise Google, Dataprovider) oder bieten auf die Wünsche von Kunden zugeschnittene Web Scraping-Produkte an (Scrapinghub, Octoparse uvm.). Beim Erwerb von gescrapten und (zum Teil) aufbereiteten Daten sind häufig jedoch nicht alle Schritte der Datengewinnung und -aufbereitung für das Statistikamt transparent. Deswegen ist es teilweise nicht möglich, die Einhaltung der hohen Qualitätsanforderungen der amtlichen Statistik zu überprüfen. Da die Transparenz verwendeter Methoden für die amtliche Statistik eine große Rolle spielt[8], stellt dies ein besonderes Problem dar.

Diese Problematik besteht jedoch für viele neue digitale Datenquellen und wird von Eurostat mit dem Konzept der Trusted Smart Statistics adressiert (Ricciato et al. 2019). Benötigte Rechenschritte von den Rohdaten bis hin zum gewünschten Output werden dabei flexibel zwischen den Datenbesitzern und dem Statistikamt aufgeteilt. Zwei Extremfälle sind denkbar: einerseits könnte das Statistikamt die Rohdaten erwerben (beziehungsweise durch Web Scraping erheben) und alle Zwischenschritte selber ausführen, wie dies in den bisherigen Web Scraping-Anwendungen in der deutschen amtlichen Statistik geschehen ist. Andererseits ist es denkbar, dass die Datenbesitzer die Daten bereits in die gewünschte Output-Form bringen und nur diesen Output an das Statistikamt liefern. Zwischen diesen beiden Extremen ist es denkbar, dass Datenbesitzer bestimmte Schritte der Datenaufbereitung und -Aggregation durchführen und dem Statistikamt ein Zwischenprodukt liefern. Die Ausführung der Programme, die einen bestimmten Output generieren, sollte getrennt werden von der Entwicklung dieser Programme. Letzteres sollte weiterhin in der Verantwortung des Statistikamtes durchgeführt werden, um die Einhaltung von Qualitäts- und Datenschutzprinzipien zu sichern und transparent zu machen. Trusted Smart Statistics setzen deswegen auch weiterhin eine hohe methodische Kompetenz der Mitarbeitenden des Statistikamtes für die jeweilige Datenquelle voraus. Mit diesem Konzept könnten technologische Hürden in den Statistikämtern umgangen werden und auf die bald jahrzehntelange Erfahrung einiger kommerzieller Unternehmen mit Web Scraping aufgebaut werden. Um dieses Konzept für Web-Daten zu verwirklichen, arbeitet Eurostat momentan an der Einrichtung eines Web Intelligence Hub, das eine einheitliche IT-Infrastruktur für Web Scraping-Aktivitäten im gesamten ESS bereitstellen soll, wie in DIME/ITDG SG (2020) beschrieben wird. Im Web Intelligence Hub als Teil des neu einzurichtenden Trusted Smart Statistics Center sollen Software-Komponenten zur Gewinnung und Analyse von Daten aus dem Web sowie ausreichende Rechenkapazitäten für die Mitglieder des ESS zur Verfügung gestellt werden. Die Funktionsweise des Web Intelligence Hub soll zu Beginn an den Anwendungsgebieten Online-Stellenanzeigen und Web Scraping von Unternehmensdaten geprüft werden, um es dann auch auf weitere Anwendungsgebiete ausweiten zu können.

In diesem Beitrag wurden zahlreiche Möglichkeiten zur Nutzbarmachung von Daten aus dem Web für die amtliche Statistik in Deutschland aufgezeigt: die eigene Erhebung von Daten durch Web Scraping, die Nutzung einer von Webseitenbetreiber-

[8] Siehe den Verhaltenskodex für europäische Statistiken (https://ec.europa.eu/eurostat/de/web/products-catalogues/-/ks-02-18-142).

Innen angebotenen API, die Nutzung des von Eurostat geplanten Web Intelligence Hub, das sich momentan in der Entwicklung befindet, und der Erwerb von gescrapten Daten von externen Organisationen, insbesondere kommerzieller Unternehmen. Welche dieser Zugangswege genutzt werden können und sollten, muss für jedes Anwendungsgebiet individuell entschieden werden.

Danksagung Mein Dank geht an Normen Peters, Jörg Feuerhake, Christian Blaudow, Jakob de Lazzer und Wolfhard Kaus für hilfreiche Kommentare und Anregungen zu diesem Beitrag.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Australian Government (2020) Vacancy report. Labour market information portal. http://lmip.gov.au/default.aspx?LMIP/GainInsights/VacancyReport. Zugegriffen: 30. März 2020

Azar JA, Marinescu I, Steinbaum MI, Taska B (2018) Concentration in US labor markets: evidence from online vacancy data. Working Paper Nr. 24395. National Bureau of Economic Research. https://doi.org/10.3386/w24395

Barcaroli G, Scannapieco M, Summa D (2016) On the use of Internet as a data source for official statistics: a strategy for identifying enterprises on the web. Italian Rev Econ Demogr Stat 70(4):25–41

Beręsewicz M, Lehtonen R, Reis F, Di Consiglio L, Karlberg M (2018) An overview of methods for treating selectivity in big data sources. Eurostat Statistical Working Paper. https://doi.org/10.2785/312232

Blaudow C, Burg F (2018) Dynamische Preissetzung als Herausforderung für die Verbraucherpreisstatistik. WISTA 2/2018:11–22

Blaudow C, Ostermann H (2020) Entwicklung eines generischen Programms für die Nutzung von Web Scraping in der Verbraucherpreisstatistik. WISTA 5/2020:103–113

Brenzel H, Czepek J, Kubis A, Moczall A, Rebien M, Röttger C, Szameitat J, Warning A, Weber E (2016) Neueinstellungen im Jahr 2015: Stellen werden häufig über persönliche Kontakte besetzt (IAB-Kurzbericht, S. 6). Institut für Arbeitsmarkt- und Berufsforschung. http://doku.iab.de/kurzber/2016/kb0416.pdf. Zugegriffen: 25. März 2020

Condron A, Kowarik A, Summa D, Stateva G, Maslankowski J, ten Bosch O et al (2019) ESS web-scraping policy template. Deliverable C1 of ESSnet Big Data II.

Daas PJH, Puts MJH (2014) Social media sentiment and consumer confidence. ECB Statistics Paper (No. 5). https://doi.org/10.2866/11606

Dautzenberg K, Gaßmann C, Groß B, Müller F, Neukamp D et al (2018) Individualisierte Preisdifferenzierung im Online-Handel. Verbraucherzentrale Brandenburg e. V. https://www.verbraucherzentrale.de/sites/default/files/2019-09/marktwaechter-untersuchung-individualisierte-preisdifferenzierung.pdf. Zugegriffen: 27. Okt. 2020

Deutscher Bundesrat (2019) BR-Drucksache 402/19 vom 30.08.2019: Entwurf eines Gesetzes zur Änderung des Gesetzes über die Preisstatistik

Directors of Methodology and IT Directors Group - Steering Group [DIME/ITDG SG] (2020) Item 3 of the agenda: Web Intelligence Hub. Meeting 12 of February 2020. https://ec.europa.eu/eurostat/cros/system/files/03_-_web_intelligence_hub.doc. Zugegriffen: 31. März 2020
Gentzkow M, Kelly BT, Taddy M (2019) Text as data. J Econ Lit 57(3):535–574. https://doi.org/10.1257/jel.20181020
Hackl P (2016) Big data: what can official statistics expect? SJI 32(1):43–52. https://doi.org/10.3233/SJI-160965
Hansen M (2020a) Dynamische Preissetzung im Onlinehandel: zu den Auswirkungen auf den Verbraucherpreisindex. WISTA 5/2020:91–102
Hansen M (2020b) Dynamische Preissetzung im Onlinehandel: zur langfristigen Anwendung von automatisierter Preiserhebung. WISTA 3/2020:14–23
Kaus W (2019a) Auslandsverlagerung wirtschaftlicher Aktivitäten: Unternehmenscharakteristika und Beschäftigungswirkung. WISTA 3/2019:11–24
Kaus W (2019b) Organisation und Verlagerung wirtschaftlicher Aktivitäten—Methodische Erläuterungen und Ergebnisse 2016. Statistisches Bundesamt. https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Unternehmen/Publikationen/Downloads-Wirtschaftliche-Aktivitaeten/verlagerung-aktivitaeten-5529301169004.pdf?__blob=publicationFile. Zugegriffen: 25. März 2020
Kinne J, Krüger M, Lenz D, Licht G, Winker P (2020) Corona-Pandemie betrifft Unternehmen unterschiedlich. Tagesaktuelle Webseiten-Analyse zur Reaktion von Unternehmen auf die Corona-Pandemie in Deutschland. ZEW-Kurzexpertise 20-05. https://www.zew.de/fileadmin/FTP/ZEWKurzexpertisen/ZEW_Kurzexpertise2005.pdf. Zugegriffen: 8. Juni 2020
Kosala R, Blockeel H (2000) Web mining research: a survey. SIGKDD Explor 2(1):1–15
Kühnemann H, van Delden A, Windmeijer D (2020) Exploring a knowledge-based approach to predicting NACE codes of enterprises based on web page texts. Stat J IAOS 36(3):807–821. https://doi.org/10.3233/SJI-200675
Mitchell R (2018) Web scraping with Python: collecting more data from the modern web, 2. Aufl. O'Reilly Media, Sebastopol
Oostrom L, Walker AN, Staats B, Slootbeek-Van Laar M, Azurduy SO, Rooijakkers B (2016) Measuring the internet economy in The Netherlands: a big data analysis. CBS Discussion Paper. https://www.nederlandict.nl/wp-content/uploads/2016/10/measuring-the-internet-economy.pdf. Zugegriffen: 25. März 2020
Peters N (2018a) Webscraping in der Beherbergungsstatistik – Ein Zwischenbericht. StaWi – Staat und Wirtschaft in Hessen Nr. 4, Hessisches Statistisches Landesamt. https://statistik.hessen.de/sites/statistik.hessen.de/files/Aufsatz_Webscraping_Beherberungsstatistik_04_18.pdf. Zugegriffen: 25. März 2020
Peters N (2018b) Webscraping von Unternehmenswebseiten und maschinelles Lernen zum Gewinnen von neuen digitalen Daten [Sonderveröffentlichung]. Hessisches Statistisches Landesamt. https://statistik.hessen.de/sites/statistik.hessen.de/files/Webscraping_von_Unternehmenswebseiten.pdf. Zugegriffen: 25. März 2020
Quaresma S, Maślankowski J, Salgado D, Ascari G, Brancato G, Di Consiglio L et al (2020) Revised version of the quality guidelines for the acquisition and usage of big data. Deliverable K3 of ESSnet big data II
Rengers M (2018a) Internetbasierte Erfassung offener Stellen im Statistischen Bundesamt. In: König C, Schröder J, Wiegand E (Hrsg) Big Data: Chancen, Risiken, Entwicklungstendenzen. Springer, Wiesbaden, S 61–86 https://doi.org/10.1007/978-3-658-20083-1_6
Rengers M (2018b) Internetgestützte Erfassung offener Stellen. WISTA 5/2018:11–33
Ricciato F, Wirthmann A, Giannakouris K, Skaliotis M (2019) Trusted smart statistics: motivations and principles. Stat J IAOS 35:1–15
Sacchi S (2014) Lange Messreihen zur Entwicklung des Stellenangebots der Schweizer Wirtschaft: Kombinierter Presse-Online-Index. SMM Working Paper 2014-1. https://doi.org/10.7892/boris.67588
von Schönfeld M (2018) Screen Scraping und Informationsfreiheit. Schriften zum geistigen Eigentum und zum Wettbewerbsrecht, Bd. 101. Nomos, Baden-Baden https://doi.org/10.5771/9783845292397-19
Stateva G, ten Bosch O, Windmeijer D, Maslankowski J, Giulio B, Scannapieco M et al (2018) Final report. Deliverable 2.4 of ESSnet big data I
Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, S 1480–1489
Zander-Hayat H, Reisch LA, Steffen C (2016) Personalisierte Preise – Eine verbraucherpolitische Einordnung. VuR 31(11):403–409