WEB SCRAPING

EINES ONLINEFORUMS FÜR DIE BINNEN-SCHIFFFAHRTSSTATISTIK – ERFAHRUNGEN UND NUTZUNGSMÖGLICHKEITEN

Alexander Brand, M.A., Dipl.Math.oec.Univ. Andreas Nickl, Felix Schmitt, Dipl.Geogr.Univ. Susanne Wilhelm

Der Einsatz von Web Scraping bei der Erstellung amtlicher Statistiken kann einen Mehrwert generieren, zum Beispiel durch eine Reduktion von Aufwand und Kosten bei der Datenerhebung, eine Entlastung der Befragten und/oder eine Verbesserung der Datenqualität. Im Rahmen einer Potenzialanalyse entwickelte ein organisationsübergreifend aufgestelltes Projektteam des Bayerischen Landesamts für Statistik einen Web-Scraping-Prozess für die Güterverkehrsstatistik in der Binnenschifffahrt, mit dem der Erhebungsablauf vereinfacht werden konnte. Mit diesem erfolgt die für die Plausibilisierung notwendige Bereitstellung von Schiffsinformationen nun nicht mehr manuell per Internetrecherche, sondern wird mithilfe einer einfachen Webanwendung automatisiert per Web Scraping durchgeführt. Inzwischen ist dieses Vorgehen als Arbeitsschritt in den Aufbereitungsprozess der Binnenschifffahrtsstatistik im Landesamt integriert worden und seit mehreren Monaten produktiv im Einsatz. Dieser Beitrag soll den Hintergrund der Projektentscheidung, den Prozess der Auswahl der genutzten Datengrundlage und die technische Implementierung vorstellen. Die vielversprechenden Resultate, welche bereits intern evaluiert wurden, lassen auf eine breitere Nutzung hoffen. Zudem wurden bereits im Aufbau Synergien für weitere Web-Scraping-Projekte im Landesamt geschaffen.

pateiflower9024 / stock.adobe.com

Bedeutung neuer digitaler Datenquellen

Wie in vielen anderen Bereichen kann die Nutzung von Web Scraping beziehungsweise des automatisierten Auslesens von Webseiten auch in der amtlichen Statistik einen potenziell großen Mehrwert bieten (siehe u. a.: Bergmann 2021; Blaudow und Ostermann 2020; Kühnemann 2021). Je nach Anwendungsbereich kann dies beispielsweise eine Reduktion des Aufwands und der Kosten bei der Datenerhebung, eine Entlastung der Befragten und/oder eine Verbesserung der Datenqualität umfassen (Kühnemann 2021). All dies sind Argumente, Web-Scraping-Verfahren in der Weiterentwicklung der amtlichen Statistikerstellung zu berücksichtigen. Im Sommer 2020 wurde daher hausintern die PG Web Scraping, eine organisationsübergreifende Projektgruppe[1] zur Potenzialanalyse von Web Scraping im Bayerischen Landesamt für Statistik, eingerichtet.

Von Automatisierungsüberlegungen zu einem konkreten Projektziel

In einem ersten prototypischen Projekt wurden Arbeitsschritte der Güterverkehrsstatistik in der Binnenschifffahrt hinsichtlich möglicher Automatisierungsmöglichkeiten durch Web Scraping geprüft. Hier konnte vor allem der Bereich der Plausibilisierungsunterstützung als Einsatzgebiet identifiziert werden. Dabei kann mithilfe des Web Scrapings der Zeitaufwand für eine manuelle Internetrecherche notwendiger Informationen verringert werden. Das Ziel des Projekts war es, diese im Internet verfügbaren Informationen in einer leicht weiterverarbeitbaren Form (beispielsweise als CSV-Datei) automatisiert und regelmäßig durch ein Scraping-Skript für die Nutzerinnen und Nutzer zur Verfügung zu stellen. Die dafür notwendigen Schritte werden in den nächsten Absätzen skizziert, wobei zunächst einige Aspekte der Binnenschifffahrtsstatistik erläutert werden.

Alexander Brand, M.A.

Alexander Brand ist seit 2022 Referent im Sachgebiet „Grundsatzfragen der Statistik, Digitalisierung, Forschungsdatenzentrum, Kompetenzzentrum Analyse" des Bayerischen Landesamts für Statistik. Dort befasst er sich mit Digitalisierung und maschinellem Lernen. Davor studierte er an der Otto-Friedrich-Universität Bamberg bis 2019 Soziologie und arbeitete danach in Bamberg und Hildesheim als wissenschaftlicher Mitarbeiter in den Bereichen Mensch-Computer-Interaktion und Computational Social Science.

Bild: privat

Dipl.Math.oec.Univ. Andreas Nickl

Andreas Nickl studierte Diplom-Wirtschaftsmathematik an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Seit 2015 ist er Referent im Sachgebiet „Grundsatzfragen der Statistik, Digitalisierung, Forschungsdatenzentrum, Kompetenzzentrum Analyse" des Bayerischen Landesamts für Statistik. Er leitet dort seit 2022 das Team „Statistische Methodik und Digitalisierung" und befasst sich vor allem mit Themen der statistischen Geheimhaltung und mathematisch-statistischen Methoden.

1 Zur Projektgruppe siehe auch Bergmann (2021).

Felix Schmitt

Felix Schmitt ist seit April 2018 im Team Verkehrsstatistiken des Bayerischen Landesamts für Statistik tätig. Basierend auf seinen täglichen Erfahrungen in der Aufbereitung und Erstellung diverser Verkehrsstatistiken beschäftigt er sich im Fachbereich insbesondere auch mit Möglichkeiten, Arbeitsprozesse zu automatisieren und deren Effizienz zu verbessern.

Bild: privat

Dipl.Geogr.Univ. Susanne Wilhelm

Susanne Wilhelm leitet seit Februar 2016 das Sachgebiet „Hochschulen, Erwachsenenbildung, Tourismus und Verkehr". Sie kam nach ihrem Studium mit dem Schwerpunkt Wirtschafts- und Sozialgeographie an das Bayerische Landesamt für Statistik und war bereits als Referentin in der Stabsstelle „Presse- und Öffentlichkeitsarbeit" sowie als Sachgebietsleiterin für die Steuer- und Krankenhausstatistiken im Einsatz.

Erhebung von Daten des Güterverkehrs in der Binnenschifffahrt

Die monatliche Erhebung zum Güterverkehr in der Binnenschifffahrt erfolgt als dezentral aufbereitete Bundesstatistik gemäß § 1 Satz 1 Nr. 1 Verkehrsstatistikgesetz (VerkStatG). Sie erfasst die Ankunft und den Abgang von Schiffen in den Binnenhäfen und an sonstigen Lösch- und Ladeplätzen einschließlich der Schiffsmerkmale sowie deren ein- und/oder ausgeladenen Güter und Containermerkmale. Der Güterumschlag ergibt sich dabei aus den Meldungen der Schiffs- sowie der Frachtführer oder Verfrachter über die Aus- und Einladungen der in den bayerischen Häfen angekommenen und abgegangenen Schiffe. Hierfür übermitteln die Binnenschiffer beziehungsweise -häfen elektronische Zählkarten, die im Landesamt plausibilisiert werden. Unter anderem wird dabei eine Bereinigung fehlender oder unplausibler Schiffsmerkmale durchgeführt. Da sowohl bayern- als auch bundesweit kein nutzbares öffentliches Schiffsverzeichnis existiert, hat der Fachbereich des Landesamts zur Vereinfachung der Plausibilisierung ursprünglich ein Schiffsverzeichnis angelegt und dieses laufend durch manuelle Internetrecherchen ergänzt.

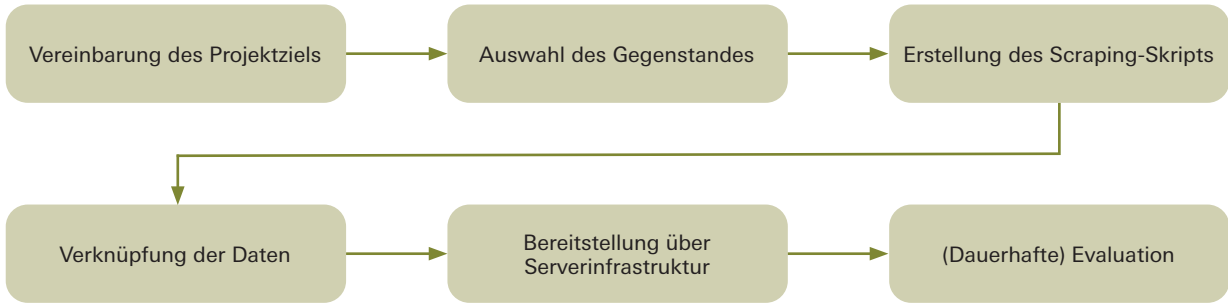
Projektablauf

Zur Durchführung eines Web-Scraping-Projektes bedarf es einiger Schritte. Der hier dargestellte Ablauf der gemeinsamen Bearbeitung im Rahmen der PG Web Scraping ist dabei als prototypisch zu verstehen. Ein derart moduliertes Vorgehen erlaubt es, die einzelnen Partizipierenden

- zuständige Fachabteilung,
- Querschnitts- und Methodiksachgebiete sowie
- IT-Bereiche

in jedem Schritt einzubeziehen, wobei der jeweils geleistete Beitrag zu den einzelnen Modulen unterschiedlich groß sein kann. So spielt die Expertise der fachlich Zuständigen bei der inhaltlichen Beurteilung möglicher auszulesender Internetseiten eine größere Rolle als beispielsweise bei der programmiertechnischen Umsetzung des Scraping-Schrittes. Die einzelnen Schritte lassen sich dabei wie folgt unterteilen (siehe Abbildung 1):

Abb. 1
Darstellung des Projektablaufs



In dem hier betrachteten Projekt mussten nach der Vereinbarung des Projektziels zunächst mögliche Informationsgrundlagen (Onlinequellen) untersucht und auf ihre Tauglichkeit für ein derartiges Vorhaben evaluiert werden (Auswahl des Gegenstandes).

Danach wurden die fachlichen Anforderungen – zum Beispiel hinsichtlich der Abrufhäufigkeit und des Datenformats der Zieldatei – komplettiert sowie technische Parameter für die Umsetzung, unter anderem die Auswahl der Programmiersprache, festgelegt. Anschließend folgte die Programmierung, die sich in drei Schritte aufgliederte: Erstellung des Scraping-Skripts, Verknüpfung der Daten und Bereitstellung über Serverinfrastruktur.

Schließlich wurde eine Evaluation des Projekts vorgenommen. Tiefer gehende Erläuterungen zu den einzelnen Punkten finden sich in den nächsten Abschnitten.

Auswahl des Gegenstandes

Für die Auswahl des Gegenstandes wurden fachseitig zunächst die wesentlichen Merkmale festgelegt, die durch das Web Scraping gewonnen werden sollten. Im nächsten Schritt wurden mehrere potenziell geeignete Internetseiten, die der Fachabteilung bekannt waren, miteinander verglichen und anhand verschiedener fachlicher und technischer Kriterien bewertet. Aus fachlicher Sicht wurden dabei vor allem die folgenden Punkte betrachtet:

- auf den Seiten verfügbarer Merkmalsumfang,
- Vorhandensein eines Identifikators für eine spätere Datenverknüpfung,
- Aktualität,
- Vollständigkeit sowie
- Datenqualität.

> *Die Einrichtung der Projektgruppe im Landesamt und die damit verbundene organisationsübergreifende Zusammenarbeit brachte Know-how zu den Methoden und der technischen Umsetzung des Web Scrapings von den IT- und Querschnittsabteilungen in die Fachabteilungen.*

Ergänzt wurden diese Kriterien durch technische Fragestellungen, wie

- der Aufwand für die Datenextrahierung,
- die Seitenhierarchie,
- die Seiten- und
- Datenstruktur (bzw. Struktur eines Eintrags),
- die Nutzungsmöglichkeit einer Suchfunktion sowie
- für das Scraping der betroffenen Seite geeignete Frameworks bzw. Programmiersprachen.

Basierend auf dieser Bewertung wurde im hier dargestellten Anwendungsfall der Binnenschifffahrtsstatistik die Internetseite www.binnenschifferforum.de (siehe Screenshot in Abbildung 2) ausgewählt.

Auf diese Internetseite wurde bereits im Rahmen der ursprünglichen Pflege des manuellen Schiffsverzeichnisses vorwiegend zurückgegriffen, da sie sich in der Plausibilisierung bereits als bewährte Hauptinformationsquelle erwiesen hat. Die eingangs erfolgte Gegenüberstellung mehrerer potenziell geeigneter Seiten bietet dennoch einen Mehrwert, da sie eine darüber hinausgehende, objektive Einschätzung aus fachlicher und technischer Sicht ermöglicht. Hier zeigt sich zudem die besondere Bedeutung der Beteiligung der fachlichen Expertise, die eine effiziente Vorselektion ermöglicht.

Die ausgewählte Datengrundlage www.binnenschifferforum.de weist alle relevanten Merkmale auf und überzeugt zudem durch eine relativ hohe Vollständigkeit: Rund 80 % bis 90 % der bisher erfassten Schiffe sind enthalten. Im Vergleich zu anderen potenziellen Quellen liegen zudem Informationen über das Datum der Einträge vor, sodass eine Einschätzung der Aktualität möglich ist. Besonders positiv ist zudem das Vorhandensein der Schiffsnummer als nutzbarem Identifikator für eine spätere Datenverknüpfung.

Die Seite weist darüber hinaus auch einige technische Eigenschaften auf, welche eine Automatisierung des Abzugs stark erleichtern. Zum einen wird eine Standard-Foren-Software genutzt (vBulletin), die einen einheitlichen Seitenquellcode erzeugt und so ein regelmäßiges Muster für den Web Scraper aufweist.

Abb. 2
Darstellung der Website mit Ansicht der genutzten Schiffsdatenbank
(www.binnenschifferforum.de/forumdisplay.php?1003-Bilder-Daten-Fakten-zu, abgerufen am 08.03.2024)

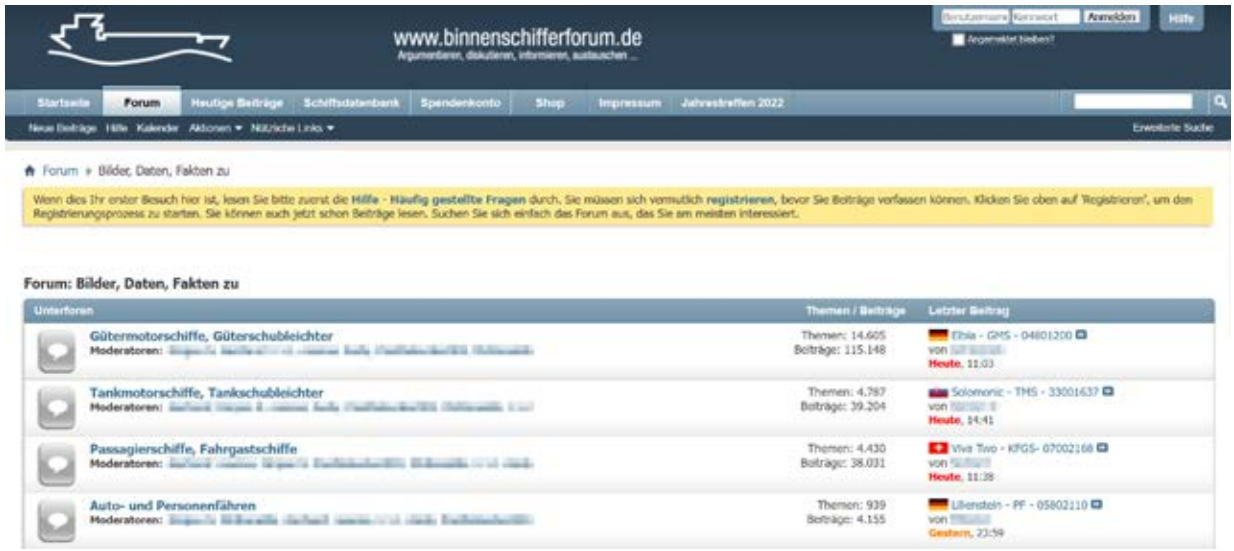


Abb. 3
Darstellung der Forenstruktur

Schematische Ablagestruktur

```
"Forum: Bilder, Daten, Fakten zu"
├─ "(Schiffskategorie-)Forum: Gütermotorschiffe, Güterschubleichter"
│   ├─ "Forum: Zahlen 0-9" (Seite 1)
│   │   ├─ "(Schiffs-)Thread: OL 3561 - GSL - 05801870" (Seite 1)
│   │   │   └─ erste Beitrag auf erster Thread-Seite enthält den Textblock "Schiffsdaten"
│   │   └─ ... 19 weitere Threads pro Seite
│   ├─ "Forum: Zahlen 0-9" (... weitere Seiten)
│   │   └─ ...
│   └─ ... weitere Foren "A" bis "Z"
└─ ... 4 weitere Unterforen zu Schiffskategorien
```

Eine API-Zugangsmöglichkeit ist hier nicht vorhanden. Obwohl diese einige Vorteile in Bezug auf die Sicherstellung der Datenstruktur mit sich bringen würde (Bergmann 2021), zeichnet sich die Seite auch ohne eine solche durch eine hohe Strukturierung aus. Die Nutzerinnen und Nutzer speichern die Schiffsinformationen per Konvention – in einer zumeist einheitlichen Form – als Forumsthread ab (der erste Beitrag im Thread enthält die Informationen als Semikolon-getrennte Schlüssel/Wert-Paare). Dies sorgt für einen relativ einheitlichen Seitenaufbau der Form Schiffsgattung, Alphabet, Schiffsname/-nummer (siehe Abbildung 3).

Erstellung des Scraping-Skripts

Für die Erstellung des Skripts zum automatisierten Abruf des Forums wurde auf die Programmiersprache Python zurückgegriffen, da diese als Lingua franca zwischen IT, Fachbereich und eher datenwissenschaftlich orientierten Mitarbeiterinnen und Mitarbeitern die Zusammenarbeit erheblich vereinfacht. Hier hat sich besonders die Notwendigkeit einer gemeinsam bearbeitbaren (Code-)Basis gezeigt, welche auch im Fall zukünftig notwendiger Änderungen einen schnellen Arbeitsablauf fördert.

Zudem erlaubt die Struktur der Seite die Nutzung eines einfach zu handhabenden Frameworks bestehend aus einer HTTP-Bibliothek (requests) und XML-Parsing (lxml). Um zusätzlich die Belastung des externen Webservers zu reduzieren und das Ausgangsmaterial für die Qualitätssicherung aufzubewahren, wird eine SQLite-Datenbank für die abgerufenen URLs genutzt.

Für die einzelnen zu extrahierenden Schiffe wird SqlAlchemy als ORM Framework verwendet, bei dem für jedes Objekt der Klasse Schiff in einer Tabelle die notwendigen Informationen (z. B. Zeitstempel, Schiffstyp und Tragfähigkeit) abgebildet werden. Im Skript werden dann alle gesammelten URLs strukturell ähnlich der Seitenhierarchie mithilfe verschachtelter Schleifen durchlaufen[2] und die Ergebnisse zuletzt aus der Datenbank in eine CSV-Datei exportiert. Die Extraktion der notwendigen Informationen wird dabei über die Pfadbeschreibungssprache XPath vorgenommen.

2 All dies sorgt für eine suboptimale Geschwindigkeit beim Seitenabruf. Das ist jedoch teils gewollt, da die Anzahl an Seitenaufrufen gesteuert werden kann. Um hier eine geringere Belastung zu erreichen, werden ebenfalls künstliche Pausen zwischen Aufrufen genutzt.

Technische Grundlagen – Glossar

Zur Erklärung der Implementierung verwendet der Beitrag einige spezifische Begriffe, welche hier in Kürze dargestellt werden. Die Reihenfolge orientiert sich im Wesentlichen am Ablauf im Skript.

1. Datenbankerstellung: **SqlAlchemy als ORM Framework**

   SqlAlchemy (www.sqlalchemy.org, abgerufen am 12.01.2024) ist ein Open-Source-SQL-Toolkit und damit ein zusätzliches Modul für Python, welches die Ablage von Informationen in einer relationalen Datenbank vereinfacht.
   SQL (Structured Query Language) ist eine Standardsprache für das Abrufen von Daten aus einer Datenbank (aws.amazon.com/de/what-is/sql/, abgerufen am 12.01.2024).
   ORM steht dabei für object-relational mapping (www.informatik-verstehen.de/lexikon/objektrelationale-abbildungen, abgerufen am 12.01.2024) zur Ablage von Informationen in einer relationalen Datenbank mithilfe einer objektorientierten Programmiersprache.

2. Iterativer Prozess des Datenbestandsaufbaus: **HTTP-Bibliothek (requests) und XML-Parsing (lxml)**

   Die requests (pypi.org/project/requests, abgerufen am 12.01.2024) und lxml (lxml.de, abgerufen am 12.01.2024) Python-Module werden genutzt, um aus dem HTML-Code der Website die entsprechenden Informationen auf Basis eines Matching-Verfahrens zu finden beziehungsweise zu analysieren.

3. Extraktion mithilfe Pfadbeschreibungssprache: **XPath**

   Entsprechende Informationen werden dann im HTML-Code mithilfe von XPath (einem System von Referenzierungsregeln für Codeelemente in strukturierten Texten) extrahiert. Die Nutzung zur Extraktion wird vom World Wide Web Consortium (W3C) empfohlen (Robie, Dyck und Spiegel, 2015).

Verknüpfung der Daten

Basis für das neu zu schaffende Schiffsverzeichnis war die bestehende, bislang manuell von der Fachabteilung erstellte Listung der Schiffe in Microsoft Excel. Das erste Scraping-Ergebnis wurde daher über die Variable ENI-Nummer[3] mit der bestehenden Liste verknüpft. In der praktischen Umsetzung zeigten sich bei der Verknüpfung jedoch Schwierigkeiten. Um verlässlich eine passende Verknüpfung mit dem ursprünglichen Schiffsverzeichnis herzustellen, mussten daher nach dem Export der Daten zunächst einige Anpassungen an der Verknüpfungsvariablen und an der Datenstruktur (Umschlüsselungen) vorgenommen werden. Hierzu wurden für die Umschlüsselung der Schiffsgattung (Codierung folgend der Spezifikation nach Datensatzbeschreibung) und der Nationalität des Schiffes zwei Hilfsdateien mithilfe der Statistiksoftware SAS an die Daten angespielt und nach der Verknüpfung mit den vorhandenen Daten als XLSX-Datei exportiert.

Entsprechend erfolgt auch die in regelmäßigen Abständen durchzuführende Aktualisierung und Ergänzung des Verzeichnisses. Im Rahmen der Plausibilisierung wird zusätzlich ein Bericht über die erstellte Datenstruktur erzeugt.

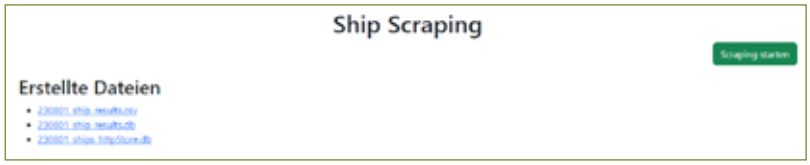
Bereitstellung über Serverinfrastruktur

Im Folgenden wird das minimalistische Interface für den Datendownload kurz vorgestellt. Um den Nutzerinnen und Nutzern (in der Fachabteilung) eine Low- beziehungsweise No-Code-Umgebung für den eigenständigen und von der Technikabteilung weitestgehend unabhängigen Datenabruf bereitzustellen, wurde eine Abrufmöglichkeit über eine Python-Anwendung (siehe Abbildung 4) mithilfe von Flask[4] erstellt. Das dargestellte Interface ist dabei nach Login für die Bearbeiterinnen und Bearbeiter zugänglich und bietet nutzerfreundliche User-Interface-Elemente, wie zum Beispiel einen Fortschrittsbalken während des Scrapings zur Überwachung, ob der Abzug korrekt angestoßen wurde.

Bezüglich des Scrapings werden zwei Abrufmöglichkeiten angeboten. Einerseits werden die Ergebnisse eines regulären monatlichen Abzugs der Website für die Bearbeiterinnen und Bearbeiter im Fachbereich zur Verfügung gestellt. Hierbei werden Abzüge ausschließlich zu unüblichen Nutzungszeiten vorgenommen, um eine Beeinträchtigung der Seitennutzungsmöglichkeiten durch Zugriffe durch den Scraper zu reduzieren. Weiterhin wurde der Seitenbetreiber über entsprechende Aktivitäten informiert[5].

Andererseits wird auch die Möglichkeit eines (zusätzlich) manuell angestoßenen Abzugs geboten. Dieser kann genutzt werden, um möglicherweise notwendige zusätzliche Datenabzüge zu testen.

Abb. 4
Interface für die Datenbereitstellung



3 Hierbei handelt es sich um eine einheitliche und eineindeutige Registrierungsnummer (European Number of Identification) für Binnenschiffe (Schiffsnummer) in Europa.

4 flask.palletsprojects.com/en/2.3.x, abgerufen am 12.01.2024.

5 Orientiert wurde sich dabei sowohl an wayback.archive-it.org/12090/20231229180654/https://cros-legacy.ec.europa.eu/content/WPC_ESSnet_Webscraping_policy_draft_en (abgerufen am 12.01.2024) als auch an statistik.hessen.de/ua (abgerufen am12.01.2024). Zusätzlich können die Statistischen Ämter nach § 5 Abs. 5 Bundesstatistikgesetz (BStatG) Daten aus „allgemein zugänglichen Quellen“ auch ohne Gesetz oder Rechtsverordnung erheben und dürfen zur Pflege und Führung des Statistikregisters gemäß § 13 Abs. 2 Satz 4 BStatG Angaben aus allgemein zugänglichen Quellen verwenden. All dies stellt die Grundlage für die Einschätzung eines rechtlich und ethisch sicheren Vorgehens im Projekt dar.

In beiden Fällen werden die Nutzerinnen und Nutzer nach erfolgtem Durchlauf des Skripts über den erfolgreichen Abzug der Daten via E-Mail informiert. Die gescrapten Daten werden jeweils als CSV-Datei zur Verfügung gestellt und können im nächsten Schritt mit dem vorhandenen SAS-Skript verknüpft und weiterverarbeitet werden.

Bei der Erstellung der Plattform hat sich gezeigt, dass Synergieeffekte genutzt werden können. So konnte das benannte Interface über eine bereits bestehende Infrastruktur für Dashboards in relativ kurzer Zeit bereitgestellt werden.

Evaluation

Das vorgestellte Projekt illustriert Möglichkeiten, wie Web Scraping in der amtlichen Statistik angewendet werden kann, um einen konkreten Mehrwert zu schaffen. Im Anschluss an die Implementierung und nach mehrfach erfolgter Bereitstellung über das Interface (siehe Abbildung 4) wurde eine erste Evaluation des Projektes vorgenommen, um eine Bewertung des Nutzens zu erhalten. Hierbei lässt sich ein positives Bild zeichnen.

Die erstmalige Entwicklung für den automatisierten Abruf benötigte zunächst einigen Aufwand mit größeren Zeitbedarfen für eine Anforderungsanalyse im Vorfeld der Programmierung sowie für die eigentliche Programmierung. Ein weiterer Entwicklungsaufwand im Bereich des Verfahrens ist nach der erstmaligen Bereitstellung jedoch nur bei Änderungen an der Website zu erwarten und dürfte voraussichtlich nur noch geringe Aufwände nach sich ziehen. Für den Betrieb des Web Scrapers wird die im Haus bereits vorhandene IT-Infrastruktur beziehungsweise Hardware verwendet. Auf diese Weise können Kosten durch die Nutzung bereits bestehender Infrastruktur gering gehalten werden.

Auf der Nutzenseite werden Rechercheprozesse vereinfacht und Effizienzgewinne in der Bearbeitungszeit erreicht. Dabei wurde die Bearbeitungszeit für die manuelle, monatliche Recherche von geschätzt bis zu vier Stunden auf aktuell eine halbe Stunde verkürzt. Allgemein ist von sich aufsummierenden positiven Effekten bei fortgeführter Nutzung auszugehen, solange keine größeren Anpassungen aufgrund von Änderungen der Seitenstruktur vorgenommen werden müssen.

Das Projekt hatte auch über den primären fachlichen Nutzen hinaus positive Effekte, die nicht direkt messbar sind. Die Einrichtung der Projektgruppe im Landesamt und die damit verbundene organisationsübergreifende Zusammenarbeit brachte Know-how zu den Methoden und der technischen Umsetzung des Web Scrapings von den IT- und Querschnittsabteilungen in die Fachabteilungen. Gleichzeitig wurden die fachlichen Besonderheiten und Bedürfnisse über die Sachgebietsgrenzen hinaus transportiert und ein gemeinsames Verständnis für Abläufe der Statistikerstellung gefördert. Nicht zuletzt partizipierten alle Beteiligten an den recherchierten Informationen zu rechtlichen sowie ethischen Gesichtspunkten des Web Scrapings. Es konnten zudem wertvolle Erfahrungen in den Bereichen Konzeption, Stakeholdermanagement und Einbettung modellhafter Arbeiten in die regulären Arbeitsschritte der amtlichen Statistik gesammelt werden.

Fazit und Ausblick

Im Jahr 2021 wurden erstmalig Daten an das Verkehrsstatistikteam des Landesamts geliefert. Der Prozess wurde in den Jahren 2022 und 2023 weiter optimiert und automatisiert. Inzwischen können der Scraping-Abruf und die anschließende Verknüpfung in SAS selbstständig vom Verkehrsstatistikteam angestoßen werden. Aktuell wird das Web Scraping monatlich automatisiert durchgeführt. Bei gleichbleibend positiver Bewertung könnte ein Test für andere Statistische Landesämter angeboten werden, um verbundweite Vorteile zu erzielen.

Innerhalb des Landesamts sollen aufbauend auf diesem vergleichsweise kleinen Projekt die Verknüpfungsalgorithmen, angewandten Methoden sowie die eingesetzte Software stetig aktualisiert und verbessert werden. Wenn dies gelingt, könnten die erworbenen Kenntnisse auch für andere Statistiken interessant werden und sich weitere Synergieeffekte einstellen. So kann eine Ausweitung des Web Scrapings auf andere Statistiken potenziell die Kosten verringern, indem auf bereits entwickelte und existierende Methoden, Programme sowie Hardware zurückgegriffen wird. Es ist davon auszugehen, dass eine Anpassung dieser Komponenten kostengünstiger ist als eine Neuentwicklung.

Zusammengefasst zeigt das Projekt, dass Web Scraping in der amtlichen Statistik eine positive Rolle spielen kann – vor allem, wenn alle Stakeholder beteiligt werden und auf Basis rechtlicher, ethischer und technischer Grundlagen gemeinsam an der Erstellung qualitativ hochwertiger Statistiken gearbeitet wird. ■

Literatur

Bergmann, Heiko (2021): Integration neuer Datenquellen in die amtliche Statistik: Web Scraping. In: Bayern in Zahlen, 75. Jahrgang, Heft 5, Fürth 2021.

Blaudow, Christian / Ostermann, Holger (2020): Entwicklung eines generischen Programms für die Nutzung von Web Scraping in der Verbraucherpreisstatistik. In: Wirtschaft und Statistik, 5, S. 103–113.

Kühnemann, Heidi (2021): Anwendungen des Web Scraping in der amtlichen Statistik. In: AStA Wirtschafts- und Sozialstatistisches Archiv, 14 (doi.org/10.1007/s11943-021-00280-5).

Robie, Jonathan / Dyck, Michael / Spiegel, Josh (2015): XML Path Language (XPath) 3.1; World Wide Web Consortium (MIT, ERCIM, Keio, Beihang). www.w3.org/TR/xpath-31 (abgerufen am 30.08.2023).