SSOAR

Open Access Repository
www.ssoar.info

Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement; mit Gutachten "Web Scraping in der unabhängigen wissenschaftlichen Forschung"

Veröffentlichungsversion / Published Version
Gutachten / expert report

Empfohlene Zitierung / Suggested Citation:
Rat für Sozial- und Wirtschaftsdaten (RatSWD). (2019). *Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement; mit Gutachten "Web Scraping in der unabhängigen wissenschaftlichen Forschung".* (Output Series, 4 (6)). Berlin. https://doi.org/10.17620/02671.39

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-68687-5

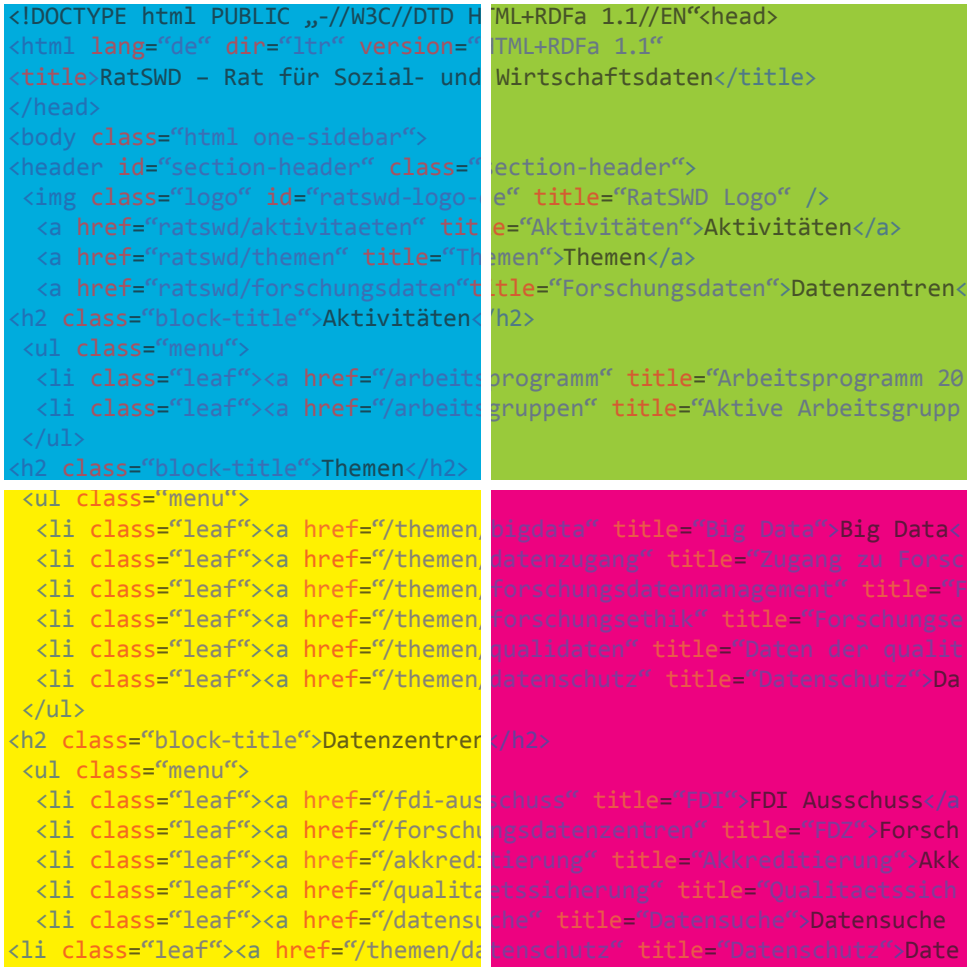
4 | Output
6. Berufungsperiode

RatSWD.
Rat für Sozial- und WirtschaftsDaten

Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften

Datenzugang und Forschungsdatenmanagement

Mit Gutachten „Web Scraping in der unabhängigen wissenschaftlichen Forschung“



GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Rat für Sozial- und Wirtschaftsdaten (RatSWD)

Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften

Datenzugang und Forschungsdatenmanagement

Mit Gutachten „Web Scraping in der unabhängigen wissenschaftlichen Forschung“

Inhaltsverzeichnis

Executive Summary 5
1 Einleitung 6
2 Zugangswege für Forschende zu Big Data 8
2.1 Individueller Zugang zu nicht-öffentlichen Daten von Unternehmen 8
2.2 Institutionalisierter Datenzugang über offene Schnittstellen 9
2.3 Sammlung von Webdaten via Screen Scraping 13
2.4 Open Data und Informationsfreiheitsgesetz 14
3 Web Scraping: Datenzugang und Forschungsdatenmanagement 17
3.1 Fallbeispiel: Reputation und Kooperation in anonymen Internetmärkten 17
3.2 Erhebung, Archivierung und Nachnutzung der erhobenen Daten 19
3.2.1 Datenzugang 19
3.2.2 Archivierung und Nachnutzung 20
4 Institutionalisierter Datenzugang über (treuhänderische) Drittpartei 21
5 Glossar 24
6 Literaturverzeichnis 26
Mitwirkende bei der Erstellung 29

Anhang: Gutachten „Web Scraping in der unabhängigen wissenschaftlichen Forschung" 31

Executive Summary

Die global gespeicherte Menge an Daten wächst exponentiell. Nach der im November 2018 veröffentlichten Studie „Data Age 2025 – The Digitization of the World“ (Reinsel/Gantz/Rydning 2018) wird ein Wachstum des globalen Datenvolumens von 33 Zettabytes (ZB) in 2018 auf 175 ZB in 2025 erwartet. In den Sozial-, Verhaltens- und Wirtschaftswissenschaften besteht ein zunehmendes Interesse an der Nutzung dieser im Rahmen der zunehmenden Digitalisierung produzierten Daten, die häufig auch als ‚Big Data‘ bezeichnet werden. Dieses Interesse begründet sich u.a. darin, dass sich diese Daten durch große Fallzahlen, die Möglichkeit der Echtzeitanalyse, nicht-reaktiven Erhebungsverfahren und die Möglichkeit der Beobachtung der Interaktion zwischen Menschen auszeichnen.

Vor diesem Hintergrund sieht der Rat für Sozial- und Wirtschaftsdaten (RatSWD) die unabweisbare Notwendigkeit, das Potenzial dieser verfügbaren Daten für die wissenschaftliche Forschung systematisch zu erschließen. Es gilt Restriktionen zu überwinden, mit welchen sich die heutige Praxis der Forschung mit Big Data-Quellen in Deutschland sowohl de jure als auch de facto konfrontiert sieht. Forschenden, denen Zugang zu Big Data-Quellen von Unternehmen gewährt wurde, wird häufig der Zugang zu unternehmensstrategisch bedeutenden Variablen oder Beobachtungen verwehrt. Sie können die verwendeten Daten oftmals nicht an andere Forschende weitergeben und unterliegen dem Risiko, dass der Datenzugang vor Abschluss eines Forschungsprojekts einseitig aufgekündigt wird. Darüber hinaus können potenzielle Interessenkonflikte oder Einschränkungen der Publikation von Forschungsergebnissen resultieren. Des Weiteren ergeben sich erhebliche neue Anforderungen bei der Aufbereitung der Daten, die bei Befragungsdaten oder administrativen Daten in dieser Form nicht auftreten. Daher sollten sich Forschende bei der Verwendung von Big Data genauestens mit dem datengenerierenden Prozess und den zur Datengewinnung verwendeten Algorithmen auseinandersetzen.

- Darüber hinaus sollten sich Nutzende von Big Data-Quellen umfassend mit den rechtlichen Bedingungen der Verwendung dieser Daten befassen. Werden derartige Daten bspw. unter Verwendung eines Application Programming Interface (API) gesammelt, müssen sich die Wissenschaftlerinnen und Wissenschaftler vor der Nutzung der Daten mit den Terms of Service für die Nutzung der API auseinandersetzen.
- Viele Forschende nutzen Web Scraping, um an Daten aus dem Internet zu gelangen. In der Erkenntnis, dass mit der Verwendung dieser Methode häufig rechtliche Unklarheiten einhergehen, hat der RatSWD ein Rechtsgutachten eingeholt. Mit diesem werden insbesondere mit Web Scraping zusammenhängende Rechtsfragen des Datenzugangs, der Datennutzung und Datenarchivierung aufgearbeitet. Der RatSWD will damit einen Beitrag leisten, die wissenschaftliche Forschung mit Big Data-Quellen für die Forschenden rechtssicherer zu machen.
- Der RatSWD empfiehlt zudem, zur Stärkung der deutschen Forschungsinfrastruktur für die Sozial-, Verhaltens- und Wirtschaftswissenschaften eine unabhängige Forschungseinrichtung zu etablieren oder eine bestehende mit dem Auftrag zu betrauen, treuhänderisch einen standardisierten Zugang zu anonymisierten (Mikro-)Daten aus öffentlichen und privaten Big Data-Quellen bereitzustellen. Die dazu erforderlichen Infrastrukturen könnten ein zentrales Thema für die sich derzeit dynamisch entwickelnde Nationale Forschungsdateninfrastruktur (NFDI) sein. Schließlich sollte Open Data insbesondere von öffentlichen Einrichtungen mit Blick auf ihre potenzielle Nutzung grundsätzlich in maschinenlesbarer Form zur Verfügung gestellt werden.

1 Einleitung

Im Zuge der Digitalisierung produzieren öffentliche Einrichtungen und insbesondere private Unternehmen ein wachsendes Volumen an Sozial-, Verhaltens- und Wirtschaftsdaten. In der Forschung besteht ein zunehmendes Interesse an der Nutzung dieser Daten. Dieses Interesse begründet sich u.a. durch:

1) Große Fallzahlen: Im Unterschied zu klassischen Erhebungen ermöglichen digital protokollierte Daten ein erheblich feineres und umfangreicheres Bild sozialer und ökonomischer Zusammenhänge.
2) Die Zeitkomponente: Das Verhalten kann in Echtzeit und über lange Zeiträume hinweg beobachtet werden.
3) Nicht-reaktive Erhebungsverfahren[1]: Der Messvorgang als solcher hat im Normalfall keinen Einfluss auf das beobachtete Verhalten.
4) Netzwerkcharakter: Es lässt sich nicht nur das individuelle Verhalten, sondern häufig auch die Interaktion zwischen Menschen beobachten.

Allerdings steht dieses Potenzial für die sozial-, verhaltens- und wirtschaftswissenschaftliche Forschung, aber auch für weitere Nutzendengruppen, wie die amtliche Statistik, nur eingeschränkt zur Verfügung. Unwägbarkeiten bezüglich Datenzugang, Datenverwertung und Datenarchivierung bilden hierbei zahlreiche Restriktionen.

Die Arbeitsgruppe Big Data des Rates für Sozial- und Wirtschaftsdaten (RatSWD) stellte sich den Fragen, ob und wie sich ein gesicherter Datenzugang für die Wissenschaft erreichen lässt, die Reproduzierbarkeit von Forschungsergebnissen gewährleistet werden kann und welche datenschutzrechtlichen und ethischen Probleme bei der Forschung mit Big Data zu berücksichtigen sind. Dieser Bericht gibt einen Überblick über fachspezifische Forschungserfahrungen, welche bislang mit Big Data gemacht wurden und wie die Bedingungen des Datenzugangs in diesen Fällen konkret aussahen. Identifiziert werden spezifische Hindernisse, die sich für Forschende in diesem Zusammenhang insbesondere im Hinblick auf den Datenzugang ergeben haben.

Die Arbeitsgruppe ging in Anlehnung an Adjerid und Kelley (2018) von einem Verständnis von Big Data als nicht zu forschungsintendierten Zwecken erzeugter, großer Menge an Daten aus, welche zeitnah zur Verfügung stehen und häufig unstrukturiert vorliegen. Massenverwaltungsdaten, wie Sozialversicherungsdaten, wurden dabei bewusst nicht berücksichtigt, da die in diesem Bericht adressierten Fragen für diesen Typ von Daten schon lange Gegenstand der wissenschaftlichen und wissenschaftspolitischen Diskussion sind.

Private Unternehmen sammeln und speichern inzwischen große Mengen an Daten, die im Normalfall nicht genuin für Forschung erhoben werden, aber für sie nutzbar gemacht werden können. Darunter fallen Daten, die täglich in großen Mengen im Produktionsprozess innerhalb von Unternehmen anfallen, im Rahmen elektronischer Kommunikation gespeichert (E-Mails, SMS , Messenger-Dienste), über Webanwendungen erzeugt (Informationen und Kommunikation in sozialen Netzwerken, Anfragen bei Suchmaschinen, Kauf- und Verkaufsverhalten auf Online-Plattformen), über individuelle Endgeräte erhoben (PCs, Smart Meter, Bewegungsdaten, Ortungsdaten, Daten aus Applikationen, Kommunikationsinformationen) oder zum Kaufverhalten von Individuen über EC-Karten, Kreditkarten oder Rabattkarten erfasst werden.

1 Nicht-reaktive Verfahren erheben Daten, die infolge von alltäglichem menschlichen Verhalten entstehen, d. h. ohne Bezug zu einer möglichen Verwendung in der Forschung. Neben digitalen Verhaltensdaten zählen z.B. auch prozessproduzierte Daten zu den nicht-reaktiven Daten.

Es liegt dabei in den meisten Fällen in der Entscheidung privater Unternehmen, Daten offen zur Verfügung zu stellen. Während Forschenden teilweise strukturierte und dedizierte Datenzugänge von Unternehmen über Verträge oder verbindliche Absprachen angeboten werden, ist ein Großteil offen verfügbarer Daten für die kommerzielle Nachnutzung durch Endkundinnen und Endkunden (v.a. über Webseiten) oder andere Drittparteien intendiert.

Eine besondere Form der Erschließung von Big Data-Quellen stellt das Web Scraping dar. Mit technischen Hilfsmitteln können hiermit Big Data-Bestände aus dem Internet für Forschungszwecke nutzbar gemacht werden. Unsicherheiten bestehen jedoch bezüglich der Bedingungen einer rechtmäßigen Datenerhebung durch Web Scraping. Da diese Methode von Forschenden bereits umfassend genutzt wird und daher besondere Relevanz hat, macht die Arbeitsgruppe mit diesem Bericht das auf ihre Initiative hin erstellte Gutachten „Web Scraping in der unabhängigen wissenschaftlichen Forschung“ verfügbar. Web Scraping umfasst dabei sowohl die Abfrage von Application Programming Interfaces (API) als auch das Auslesen von für menschliche Endnutzende bestimmten Webseiten. Letztere Variante des Web Scrapings wird zur begrifflichen Abgrenzung im Folgenden als „Screen Scraping“ bezeichnet (vgl. auch unsere Definition in Abschnitt 2.3).

Der folgende Abschnitt 2 erläutert verschiedene Zugangswege zu Big Data für Forschende. Hierbei wird zwischen einem individuellen und einem institutionellen Zugang zu Big Data sowie der Sammlung von Daten über unregulierte Zugangswege, konkret über Screen Scraping, unterschieden. Schließlich diskutiert dieser Abschnitt Möglichkeiten und Probleme der Verwendung von Open Data. Abschnitt 3 liefert eine Zusammenfassung der Erkenntnisse des auf Initiative der Arbeitsgruppe in Auftrag gegebenen rechtlichen Gutachtens zur Datensammlung über Web Scraping. Vor dem Hintergrund der aufgezeigten Probleme mit den bestehenden Zugangsmöglichkeiten wird in Abschnitt 4 die Einführung einer Datentreuhandstelle als ein mögliches Lösungsmodell vorgestellt. Abschnitt 5 enthält schließlich ein Glossar zu den verschiedenen technischen Begriffen, die in diesem Papier verwendet werden.

2 Zugangswege für Forschende zu Big Data

In diesem Kapitel werden zunächst drei in der Forschung oft genutzte Zugangswege zu Big Data skizziert. Diese Datenzugänge sind hier nach dem Maß der bewussten Einwilligung der Datenerzeugerin (zumeist ein Privatunternehmen) angeordnet. Der Zugang ist entweder durch bilaterale Vereinbarungen (individueller Datenzugang, Abschnitt 2.1), one-to-many Vereinbarungen (institutionalisierter Zugang, Abschnitt 2.2), oder ein Fehlen an Vereinbarungen mit dem Datenabgreifenden (Screen Scraping, Abschnitt 2.3) charakterisiert. Als potenzielle Alternativquellen für Big Data werden schließlich in Abschnitt 2.4 Open Data und die Informationsfreiheit vorgestellt.

2.1 Individueller Zugang zu nicht-öffentlichen Daten von Unternehmen

Viele Forschende vereinbaren mit privaten Unternehmen einen individuellen Datenzugang für spezifische Forschungsprojekte (siehe Edelman 2012 sowie Einav und Levin 2014 für eine Übersicht entsprechender Forschungsprojekte).

Beispiel 1: Kooperation des RWI – Leibniz-Institut für Wirtschaftsforschung mit ImmobilienScout24 (an de Meulen/Micheli/Schaffner 2014)

Im Rahmen dieser Kooperation stellte ImmobilienScout24.de dem RWI alle auf seiner gleichnamigen Webplattform gesammelten Informationen für wissenschaftliche Analysen zur Verfügung (siehe Boelmann und Schaffner 2019, für eine Datensatzbeschreibung). Die Daten umfassen jedes Angebotsinserat, das von 2007 bis zur aktuellen Verfügbarkeit auf der zugrundeliegenden Plattform eingestellt wurde. ImmobilienScout24.de stellt derzeit die marktführende Plattform des Immobilienmarkts im deutschsprachigen Raum dar. Neben der Aktualität und dem Informationsgehalt der verfügbaren Daten zeichnen sich diese durch ihre geografische Lokalisierbarkeit aus, da alle auf ImmobilienScout24.de angebotenen Objekte georeferenziert sind. Damit können nicht nur kleinräumige Entwicklungen beobachtet, sondern die Daten können aufgrund der Georeferenzierung auch mit einer Vielzahl von Informationen auf jedweder regionalen Aggregationsebene integriert werden.

Beispiel 2: Kooperation des Statistischen Bundesamtes mit Mobilfunkunternehmen

Das Statistische Bundesamt (destatis) kooperiert mit Unternehmen, um Zugang zu deren digitalen Daten zu erhalten. Im Rahmen einer Machbarkeitsstudie zur Erforschung der Potenziale von Mobilfunkdaten in der amtlichen Statistik ist Destatis im September 2017 eine Kooperation mit zwei Tochterunternehmen der Deutschen Telekom AG eingegangen, der T-Systems International GmbH und der Motionlogic GmbH (Hadam 2018). Ziel dieses Projekts ist zu prüfen, inwieweit mit Mobilfunkdaten die Tages- und Wohnbevölkerung, Pendlerströme sowie die Verteilung von Touristen valide abgebildet und geschätzt werden können.

Ein Datenzugang, wie in den Beispielen 1 und 2 skizziert, wird häufig nur gewährt, wenn das Unternehmen von dem spezifischen Forschungsprojekt profitiert. So können Unternehmen zusätzliche Erkenntnisse über die Potenziale der eigenen Daten oder das Verhalten ihrer Kunden gewinnen, über die Erkenntnisse des Projekts neue Produkte oder Dienstleistungen für die Kunden entwickeln oder die eigene Organisation (bspw. das interne Datenmanagement) optimieren. Aus einem derartigen (projektspezifischen) individuellen Zugang können allerdings vielfältige Probleme resultieren. Forschende sollten die entstehenden Risiken sorgfältig abwägen.

Hinweise und Empfehlungen:

- Die datengenerierenden Prozesse sind oft nicht komplett transparent. So kann das Nutzendenverhalten durch maschinell generierte Empfehlungen beeinflusst sein, welche sich über die Zeit verändern können. Dazu muss beim Datenproduzierenden konkret und schon zu Beginn der Kooperation nachgefragt werden.
- Den Forschenden werden häufig nicht alle potenziell für die Forschungsfrage relevanten Daten zur Verfügung gestellt und (unternehmensstrategisch) bedeutende Variablen oder Beobachtungen zurückgehalten.
- Zumeist muss davon ausgegangen werden, dass die Wissenschaftlerinnen und Wissenschaftler, denen Zugang zu Daten gewährt wurde, diese nicht an andere Forschende weitergeben dürfen. Damit werden die Möglichkeiten des Nachvollzugs und der Replikation von Forschungsergebnissen und damit potenziell die Optionen der Publikation der Forschungsergebnisse eingeschränkt. Auch eine Datennachnutzung wäre dann ausgeschlossen.
- Da es sich beim Datenzugang um eine vertragliche Vereinbarung handelt, kann diese in der Regel einseitig gekündigt werden. Hieraus ergibt sich ein Risiko für den erfolgreichen Abschluss von Forschungsprojekten.
- Werden die Daten den Wissenschaftlerinnen und Wissenschaftlern im Rahmen von Beratungsaufträgen zur Verfügung gestellt, ergeben sich schließlich Interessenskonflikte.

2.2 Institutionalisierter Datenzugang über offene Schnittstellen

Neben Unternehmen, die einen individuellen Datenzugang mit Forschenden vereinbaren, mehrt sich die Zahl von Unternehmen, darunter gerade Web-Plattformen, die einen Teil der von ihnen generierten digitalen Daten über öffentlich zugängliche Web-Schnittstellen (sog. Application Programming Interfaces, kurz API) für die Nutzung durch Dritte zur Verfügung stellen. Mit diesen APIs gibt das Unternehmen freiwillig einen strukturierten Einblick in – und Zugriff auf – die hinter dem jeweiligen Angebot liegenden Datenstrukturen, welcher von sehr eingeschränkt bis nahezu umfassend sein kann. Obwohl solche APIs primär für die kommerzielle Nutzung intendiert sind (z.B. in Mobil-Applikationen von Dritten), stellen sie eine vielversprechende Ressource für Forschungsvorhaben dar. Wissenschaftlerinnen und Wissenschaftler verwenden die Möglichkeiten dieser APIs daher vermehrt, um reichhaltige Daten über menschliches Verhalten, ökonomische Indikatoren und andere Sachverhalte zu gewinnen und in ihrer Forschung zu verwenden.

Die Inanspruchnahme dieser Schnittstellen erfordert die Einwilligung des Nutzenden in die Terms of Service bzw. Allgemeinen Geschäftsbedingungen des Unternehmens. Diese Einwilligung erfolgt dabei häufig implizit durch die Nutzung der Schnittstelle oder durch eine vorangehende Registrierung. Die genannten Regelungen schließen wiederum regelmäßig bestimmte Weiterverwendungen der erlangten Daten aus und erlegen teils enge Vorgaben für das Datenmanagement auf.[2]

2 Als Beispiel sei hier die Aufforderung von Twitter genannt, alle auf der Plattform gelöschten Tweets auch in nachträglich gesammelten, bei Dritten verbleibenden Datensätzen zu löschen (https://developer.twitter.com/en/developer-terms/agreement-and-policy.html, Stand 23.08.2019).

Beispiel 3:
Ein Beispiel für eine Anwendung solcher Daten ist das Angebot von “Google Trends”, welches die quantitative Entwicklung bestimmter Suchbegriffe in den Nutzendenabfragen an Googles Suchmaschine abbildet. Diese Daten werden bereits häufig für wissenschaftliche Zwecke verwendet (siehe bspw. Carrière-Swallow und Labbé 2013; Hyunyoung und Varian 2012; McLaren und Shanbhogue 2013; Bug 2015: 86; Rieckmann und Schanze 2015).

Schmidt und Vosen (2011, 2012, 2013) nutzen die in der "Insights for Search"-Applikation von Google Trends bereitgestellten Informationen zur Prognose des privaten Konsums. Die Applikation weist auf Basis von Stichproben den Anteil der Suchanfragen nach 605 Kategorien und Unterkategorien aus, die auf Wochenbasis seit dem Jahr 2004 vorliegen. Von den Autoren wurden 45 dieser Kategorien zur Schätzung eines gewichteten Konsumindikators verwendet, da diese für die privaten Konsumausgaben relevant und mit den Komponenten des privaten Konsums nach der volkswirtschaftlichen Gesamtrechnung des Statistischen Bundesamts kompatibel sind. Dabei ergab sich das Problem, dass bei einer Anfrage an die "Insights for Search"-Applikation die zugrundeliegende Software für einen bestimmten Tag eine zufällige Stichprobe aller von Google registrierten Suchanfragen generiert. Anfragen, die am selben Tag an Google Trends gestellt wurden, lieferten dieselben Ergebnisse. Jedoch variierten die Ergebnisse an verschiedenen Tagen aufgrund der unterschiedlichen von Google Trends gezogenen Zufallsstichproben aller Suchanfragen stark. Die Forschenden lösten dieses Problem, indem Durchschnitte von 52 an verschiedenen Tagen gezogenen Stichproben verwendet wurden.

Beispiel 4:
Mit dem aus dieser Forschung resultierenden „RWI-Konsumindikator“ kann die aktuelle Entwicklung des Konsums gut abgebildet werden. Schmidt und Vosen (2012) zeigen, dass mit Hilfe der „Insights for Search“-Applikation von Google Trends über die Möglichkeit der Berücksichtigung spezieller politischer Maßnahmen, wie bspw. die in vielen Ländern während der Rezession im Jahr 2009 eingeführten Abwrackprämien, Konjunkturprognosen, verbessert werden können.

Neben den rechtlichen Unwägbarkeiten weisen die Beispiele 3 und 4 darauf hin, dass auch die Erzeugung der abgerufenen Daten stets einer gründlichen Prüfung unterzogen werden müssen; dies gilt für jegliche APIs. Mögliche Probleme, die durch diese Daten entstehen können, veranschaulicht bspw. der ehemalige Service ‚Google Flu Trends‘, den Google als Werkzeug zur Grippewelle-Vorhersage entwickelt und veröffentlicht hatte. Die erzielte Voraussagegenauigkeit war in anfänglichen Evaluationen zunächst sehr hoch. Inzwischen wurde festgestellt, dass aufgrund von Problemen in der Genauigkeit und vor allem Transparenz der Daten sowie der zugrundeliegenden Messkonstrukte und vergleichsweise einfachen Prognosemodelle Google Flu Trends keine konstant verlässlichen Vorhersagen erzeugte (Lazer et al. 2014). Solch mangelnde Reliabilität, gerade über die Zeit hinweg, ist für datennutzende Forschende in den seltensten Fällen von Beginn an transparent – was die Qualität der Ergebnisse (insbesondere wenn nur auf eine einzelne, dynamische Datenquelle fußend) schlecht abschätzbar macht (Di Bella/Leporatti/Maggino 2018).[3]

3 Mit methodisch breiterem Ansatz geht z.B. das Robert Koch Institut vor: https://www.rki.de/DE/Content/Infekt/IfSG/Signale/Projekte/Forschungsantrag_Demis_Signale.pdf?__blob=publicationFile

Auch die APIs von Twitter[4] werden inzwischen häufig für empirische Forschungsprojekte verwendet (für einen Überblick siehe bspw. McCormick et al. 2017). In Forschungsdisziplinen, die mit der Informatik verwandt sind, wie Computational Social Science, gehören Twitterdaten mittlerweile sogar zum „de facto core dataset" (Pfeffer/Mayer/Morstatter 2018). Über die API von Twitter kann eine 1 %-Stichprobe aller Tweets kostenlos abgerufen werden. Eine größere Stichprobe der Tweets ist kostenpflichtig. Ausgehend von der Annahme, dass diese Stichprobe eine zufällige Auswahl vis-à-vis der Gesamtinhalte der Plattform darstellt, haben viele Forschungsprojekte bislang naturgemäß auf den kostenlosen 1 %-Zugang zurückgegriffen. Zum einen muss dabei die inhärente Nichtrepräsentativität der Twitter-Nutzendenschaft für Schlüsse auf größere (Offline-)Gesamtpopulationen beachtet werden.[5] Zusätzlich zeigen neuere Erkenntnisse auf Basis von Reverse-Engineering, dass die kostenlose 1 %-Stichprobe nicht vollkommen zufällig aus der Gesamtheit der Plattforminhalte gezogen wird und manipulierbar ist (Morstatter et al. 2013, Pfeffer/Mayer/Morstatter 2018). Vielfältige weitere Fehler in der Operationalisierung von Messkonstrukten und der Inferenz auf eine Zielpopulation können mit solch limitierten Daten auftreten und müssen im Forschungsdesign mitbedacht werden (Sen et al. 2019).

Neben Twitter stellt Wikipedia die seinem Serviceangebot zugrundeliegenden Daten öffentlich über APIs zur Verfügung. Über diese können bspw. Besuchendenzahlen für einzelne Seiten bzw. Themen sowie Textänderung erfasst werden. Auch diese Daten werden zunehmend in der Forschung verwendet (siehe bspw. Moat et al. 2013; Slivko 2018).

Vergleichbare APIs werden von vielen Webplattformen und Webserviceprovidern angeboten, von Youtube über Instagram hin zu Spezialportalen wie Stackoverflow.[6] Daneben existieren viele APIs, mit denen Mobilitätsmuster analysiert werden können. Beispiele hierfür sind die APIs von Unternehmen, die Fortbewegungsmittel verleihen.[7] Insbesondere zu Leihfahrrädern existiert inzwischen eine umfangreiche Literatur (siehe bspw. Fishman 2016, Yongping und Zhifu 2018).

Ähnlich zu Zugängen über APIs verhalten sich von Unternehmen zur Verfügung gestellte statische Datendownloads, häufig bestehend aus einer Datei-Sammlung von exportierten Datenbanken oder Datenbankteilen (sog. Data Base Dumps). Auch die Nutzung solcher herunterladbarer Daten richtet sich – wie bei Schnittstellen – nach der Einwilligung in die Nutzungs- und Copyrightbedingungen des Anbietenden. Als Beispiele seien hier die Wikimedia Foundation genannt, welche umfängliche Data Base Dumps frei zur Verfügung stellt[8], die Internet Movie Database[9] sowie das Statistikportal FiveThirtyEight[10].

4 https://developer.twitter.com/en/docs/api-reference-index.html
5 Generell sind junge, männliche Nutzer überrepräsentiert. Dies ist auch in Deutschland der Fall. Hierzulande nutzen nur knapp 4 % der Bevölkerung Twitter mindestens einmal wöchentlich (Frees und Koch 2018), wobei Nutzung häufig passive Rezeption ohne Autorenschaft oder Interaktionen meint.
6 Eine Diskussionsplattform zur Softwareentwicklung.
7 Siehe bspw. https://github.com/ubahnverleih/WoBike oder auch https://github.com/CityOfLosAngeles/mobility-data-specification/blob/dev/agency/README.md#vehicle-events
8 https://dumps.wikimedia.org
9 https://www.imdb.com/interfaces
10 https://data.fivethirtyeight.com

Hinweise und Empfehlungen:

- Forschenden ist anzuraten, die Terms of Service bzw. Nutzungsvereinbarungen oder AGB der verwendeten API detailliert zu begutachten und möglichst zu dokumentieren. Insbesondere haben einige Unternehmen Ausnahmen für wissenschaftlich Forschende in ihre Nutzungsbedingungen aufgenommen, die größere Freiheiten für Datenabruf und -nutzung erlauben.
- Die Bestimmungen der Unternehmen enden meist nicht beim Zugriff auf die Daten, sondern sehen Begrenzungen hinsichtlich der weiteren Speicherung, Weitergabe und Veröffentlichung vor. Potenzielle Konflikte mit guter wissenschaftlicher Praxis[11], gerade Reproduzierbarkeit, sind fallweise abzuwägen.[12]
- Die Einhaltung der Nutzungsvereinbarungen befreit nicht von weiteren rechtlichen und forschungsethischen Limitationen, wie z.B. der Wahrung der Datenschutzrechte der Personen, deren Verhaltensdaten die originäre Basis der nachgenutzten Datensätze darstellen, auch über die von Anbietenden gemachten Beschränkungen hinaus. Weiterhin können das Urheberrecht und selten auch das Wettbewerbsrecht betroffen sein (siehe hierzu auch Kapitel 3.2).
- Wissenschaftlerinnen und Wissenschaftler, die über APIs (oder Dateidownloads) auf Daten zugreifen, sollten sich genauestens mit dem datengenerierenden Prozess und den zur Datengewinnung verwendeten Algorithmen auseinandersetzen, insofern diese dokumentiert oder anderweitig ermittelbar sind. Besonders zu beachten: Datengenerierungsprozesse können sich schnell ändern (z.B. durch Umstellung der Plattform-Software). Validität und Reliabilität von aus solchen Daten gewonnenen Metriken sind in diesem Kontext kritisch zu prüfen.

11 Siehe bspw. „Leitlinien zur Sicherung guter wissenschaftlicher Praxis" der Deutschen Forschungsgemeinschaft (DFG): https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp.

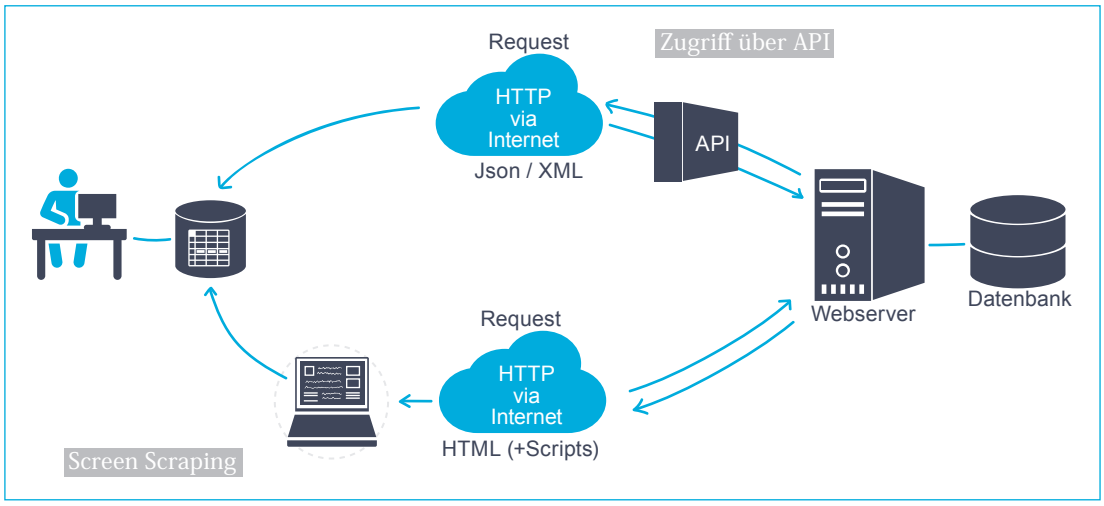
12 Auch können Konflikte der Nutzungsbedingungen des ursprünglichen Datenanbietenden mit Nutzungsbedingungen von Archivierungsplattformen (Stichwort: Permanenz von Ressourcen) oder mit nationalen Rechtsvorschriften auftreten.

2.3 Sammlung von Webdaten via Screen Scraping

In vielen Fällen sind von einer im Web vertretenen Entität weder dedizierte Zugänge zum Abrufen oder Herunterladen ihrer Daten vorgesehen, noch sind bilaterale Vereinbarungen über einen Zugriff möglich. Die von einer Web-Plattform für menschliche Nutzende über Webseiten bereitgestellten Daten sind zudem nicht immer in ihrer Gänze über spezielle Schnittstellen oder Downloads verfügbar.

In diesen Fällen greifen Forschende häufig auf sogenanntes ‚Screen Scraping' zurück. Während in der Praxis unterschiedliche Definitionen zu den Begriffen ‚Web Scraping' und ‚Screen Scraping' existieren – und beide zuweilen gleichgesetzt werden – folgt der Bericht im Weiteren der Definition von Screen Scraping als „das automatisierte Abrufen und Auslesen von Informationen und Daten aus dem unstrukturierten Teil des Internets, der für den menschlichen Nutzer bestimmt ist." (von Schönfeld 2018: 20; Hervorhebung unsere). Entsprechend simuliert „ein Screen Scraping-Programm [...] menschliches Nutzerverhalten, um Zugang zu Webseiten zu erhalten und die dort vorhandenen Informationen und Daten abzurufen und anschließend auszuwerten." (von Schönfeld 2018: 58).[13] Demgegenüber wird Web Scraping als breiterer „Oberbegriff für ein algorithmus-basiertes Verfahren [verstanden], mit dem im World Wide Web technisch frei zugängliche Informationen und Daten ausgelesen werden." (ebd.) Dies schließt neben Screen Scraping weitere Verfahren (besonders Zugriffe über Programmierschnittstellen, siehe Abschnitt 2.2) ein. Die Unterschiede zwischen Screen Scraping und den Zugriff über programmatische Schnittstellen sind in Abbildung 1 dargestellt.

Abb. 1: Zwei Arten des Web Scrapings im Vergleich: Datenzugriff über API und Screen Scraping



© RatSWD 2019

Üblicherweise wird beim Screen Scraping über eine Applikation bzw. ein Skript eine Anfrage zu einer oder mehreren vordefinierten URLs[14] an einen Server gestellt. Die mitgesendeten Metadaten emulieren dabei eine typische Browser-Applikation, wie z.B. Mozilla Firefox. Das zurückgesendete statische Dokument ist meist in HTML und CSS verfasst, in manchen Fällen auch in menschenlesbaren Dateiformaten wie PDF. Aus diesem werden solche Teile durch selbst-definierte Heuristiken ausgelesen, die für den Anfragenden von Interesse sind.[15] In einem weiteren Schritt können zudem verknüpfte Bilder und andere Multimedia-Inhalte geladen und gespeichert werden.

13 Ausgenommen davon sind die typischerweise höheren Zugriffsraten automatisierter Verfahren, welche jene von menschlichen Akteuren naturgemäß deutlich übertreffen können.

14 URLs für die Abfrage können dabei als explizit ausformulierte URL-Adresslisten vorliegen oder bloße ungefähre Muster zur Abfrage beinhalten, z.B. alle Unterseiten einer definierten Domain oder alle Webseiten mit einem bestimmten Suchbegriff im Titel.

15 Ein beispielhafter Fall: Mit dem Begriff ‚Price' wird in der HTML-Baumstruktur einer Online-Shop-Seite nach den HTML-Elementen gesucht, welche die Preise aufgelisteter Produkte enthalten, mit ‚Description' nach deren Kurzbeschreibungen. Diese Elemente werden dann als Variablen in ein strukturiertes Format überführt.

Können Inhalte nicht durch bloße URL-Abrufe erreicht werden, kommen auch Techniken zum Einsatz, die nach dem Laden einer Seite Feldeingaben oder die Bedienung von Schaltflächen über Mausklicks durch einen menschlichen Nutzenden simulieren, um weitere Informationen anzufordern. Durch dynamisches Nachladen von Seitenelementen können so z.B. Inhalte abgefragt werden, die nur durch spezielle Sequenzen von Nutzendenverhalten entstehen. Über Feldeingaben (teils auch über URL-Direktaufruf) kann die Screen Scraping-Applikation sich bei vorhandenen Login-Daten als User anmelden und so Zugriff auf von der abzufragenden Webseite nicht als ‚öffentlich' vorgesehenen Daten erlangen. Damit (und mit der vorhergehenden Registrierung) einher geht im Regelfall die Anerkennung der Nutzungsbedingungen des Dienstes – entweder implizit oder über explizite Feldeingaben. Über interaktive Elemente, über das Folgen von Hyperlinks (auch ‚Crawling' oder ‚Spidering' genannt) oder über die Erzeugung von URL-Varianten können große Teile des Gesamtangebots einer Webplattform exploriert und abgerufen werden, ohne konkrete URL- bzw. Seitenlisten als Startpunkt zu benötigen.

Viele Forschungsprojekte in den Sozial-, Verhaltens- und Wirtschaftswissenschaften setzen mittlerweile Screen Scraping-Methoden ein. So ist dies z. B. in der Produkt- und Preisforschung häufig notwendig, da offene Schnittstellen von relevanten Web-Plattformen nur in seltenen Fällen angeboten werden. Solche Forschung umfasst unter anderem Preisentwicklungen auf Portalen für Hotel- und Ferienwohnungsvermittlungen (Gyódi 2017), Mietpreisentwicklungen in öffentlichen Online-Kleinanzeigenportalen (Boeing und Waddell 2016), algorithmische Preissetzung auf Shoppingportalen (Chen/Mislove/Wilson 2016) oder Vorhersagen von Preisindizes für Konsumgüter über Webseiten großer Einzelhandelsketten (Powell et al. 2017).

Web Scraping – und insbesondere Screen Scraping – ist in vielen weiteren Forschungszweigen populär. Unter anderem untersucht wurden:

- Ethnische Segregationen durch Scraping von Reviews auf Restaurantbewertungsseiten (Davis et al. 2019),
- Varianzen in Qualifikationsanforderungen auf Jobportalen (Verma et al. 2019)
- Ungleichbehandlungen auf Freelance-Onlinemarktplätzen (Hannak et al. 2017)
- Dynamiken von politischer Berichterstattung und Konsumption im Web (Ørmen 2019)
- Neue Indikatoren, bspw. für Innovation (Kinne und Axenbeck 2018)

In der Forschungspraxis kommt dem Web Scraping heute eine besondere Bedeutung zu. Eine ausführliche Darstellung der Forschungsrelevanz anhand eines konkreten Fallbeispiels und die wesentlichen Inhalte des vom RatSWD eingeholten Rechtsgutachtens einschließlich resultierender Handlungsempfehlungen erfolgt daher gesondert in Kapitel 3.

2.4 Open Data und Informationsfreiheitsgesetz

Neben den bereits diskutierten Zugangswegen zu Big Data ermöglichen Open Data und das Informationsfreiheitsgesetz einen potenziellen Zugang zu Politik- und Verwaltungsdaten. Die zugänglich gemachten Daten wurden, wie auch bei Big Data, in der Regel nicht für Forschungszwecke erhoben. Es handelt sich dabei nicht notwendigerweise um eine große Datenmenge. Bei der Nutzung für die wissenschaftliche Forschung sind aber ähnlich wie bei den diskutierten Datenzugangswegen jeweils besondere Einschränkungen zu beachten.

Open Data

Der Begriff „Open Data" beschreibt ein Konzept, bei dem maschinenlesbare und strukturierte Informationen durch die Verwendung offener Nutzungsrechte von jedermann frei verwendet, nachgenutzt und verbreitet werden können. Die Nutzung dieser offenen Daten darf laut der Open Data-Definition nur eingeschränkt werden, um den Ursprung durch Quellennennung und die Offenheit der in ihnen enthaltenen Informationen transparent zu machen. Diese offenen Daten dürfen keine personenbezogenen Daten oder Daten, die dem Datenschutz unterliegen, beinhalten.

Im Gegensatz zu bereits verarbeiteten und meist rechtlich geschützten Informationen handelt es sich bei Open Data oft nicht nur um Text- oder Bildmaterial, sondern um Tabellen, Karten oder Datenbanken. In diesem Zusammenhang wird auch von „Rohdaten" gesprochen, die als Grundlage für die letztendlich aufbereitete Information dienen. Diese Daten können aus den unterschiedlichsten Bereichen der Gesellschaft stammen: Geodaten, Kulturdaten, Daten aus Wissenschaft und Forschung sowie Wetter- und Umweltdaten.

Definitionen

Open Data

„Offene Daten sind sämtliche Datenbestände, die im Interesse der Allgemeinheit ohne jedwede Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterverwendung zugänglich gemacht werden. Zu denken wäre etwa an Lehrmaterial, Geodaten, Statistiken, Verkehrsinformationen, wissenschaftliche Publikationen, medizinische Forschungsergebnisse oder Hörfunk- und Fernsehsendungen. Bei ‚Open Data' handelt es sich nicht ausschließlich um Datenbestände der öffentlichen Verwaltung, denn auch privatwirtschaftlich agierende Unternehmen, Hochschulen und Rundfunksender sowie Non-Profit-Einrichtungen produzieren entsprechende Beiträge." (von Lucke und Geiger 2010: 3). „Zur Kennzeichnung und Regelung der freien Nutzbarkeit von Daten dienen geeignete Lizenzen." (ebd.).

Linked Open Data

„Offene vernetzte Daten sind sämtliche Datenbestände, die im Interesse der Allgemeinheit der Gesellschaft ohne jedwede Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterverwendung frei zugänglich gemacht und über das World Wide Web miteinander vernetzt sind." (von Lucke und Geiger 2010: 4). „Mehrwerte ergeben sich, wenn Datenbestände, die zuvor noch nicht miteinander verknüpft waren, miteinander kombiniert werden und dies zu neuen Erkenntnissen führt. Vor allem die leichte Adressierbarkeit von Datenbeständen im Internet hilft, vorhandene Hürden beim Datenabruf zu senken. Mit Unterstützung von Uniform Ressource Identifiers (URI) und des Ressource Description Frameworks (RDF) lassen sich Teile von Daten, Informationen und Wissen aufbereiten, teilen, exportieren und vernetzen." (ebd.: 3).

Informationsfreiheitsgesetz (IFG)

§

Informationsfreiheit (auch Informationszugangsfreiheit, Informationstransparenz, englisch Freedom of Information (FOI)) ist ein allgemeines Grundrecht zur öffentlichen Einsicht in Dokumente und Akten der öffentlichen Verwaltung. Es kann sowohl von einzelnen Bürgerinnen und Bürgern, aber auch im Rahmen von journalistischen Recherchen und wissenschaftlichen Forschungsprojekten wahrgenommen werden. Hierfür können z. B. Ämter und Behörden verpflichtet werden, ihre Akten und Vorgänge zu veröffentlichen (Öffentlichkeitsprinzip) bzw. zugänglich zu gestalten (Verwaltungstransparenz) und zu diesem Zweck verbindliche Qualitätsstandards für den Zugang zu definieren.

Das Gesetz zur Regelung des Zugangs zu Informationen des Bundes, kurz auch Informationsfreiheitsgesetz (IFG/IFG-Bund) ist ein deutsches Gesetz zur Informationsfreiheit. Bisher haben dreizehn Bundesländer für ihren Zuständigkeitsbereich jeweils eigene ähnliche Gesetze erlassen. In Bayern, Niedersachsen und Sachsen existiert hingegen keine Landes-Informationsfreiheitsgesetze.

Das IFG enthält zahlreiche Ausnahmetatbestände, durch die das Recht auf Informationszugang eingeschränkt oder ganz verwehrt werden kann:

- Die Informationsfreiheit bezieht sich ausschließlich auf abgeschlossene dokumentierte Vorgänge, öffnet also keinen Zugang zu laufenden Planungen (§ 3 Schutz von besonderen öffentlichen Belangen, § 4 Schutz des behördlichen Entscheidungsprozesses).
- Die Informationsfreiheit schließt personenbezogene Daten (§ 5) und betriebsbezogene Daten (§ 6) aus. So darf ein Zugang zu personenbezogenen Daten nur dann gewährt werden, wenn das Informationsinteresse des Antragstellers das schutzwürdige Interesse der oder des Betroffenen überwiegt oder die/der Betroffene eingewilligt hat. Bezüglich der Inhalte von Personalakten und Personalverwaltungssystemen besteht kein Informationszugangsanspruch. Informationen über Namen und dienstliche Anschriften von Beschäftigten sollen jedoch grundsätzlich zugänglich gemacht werden. Dasselbe gilt für Informationen zu Gutachterinnen und Gutachtern und Sachverständigen.

Antragsberechtigt ist jede natürliche Person und jede juristische Person des Privatrechts (beispielsweise eingetragene Vereine). Bürgerinitiativen und Verbände, die selbst nicht als juristische Person des Privatrechts auftreten, sind nicht antragsberechtigt.
Grundsätzlich muss ein Antrag nicht begründet werden. Ausnahmen bestehen nur, wenn die Rechte Dritter betroffen sind, wenn es um den Schutz geistigen Eigentums oder von Geschäftsgeheimnissen geht. Eine Begründung ist deswegen erforderlich, damit die oder der Dritte, die oder der von der Behörde in diesen Fällen in Kenntnis gesetzt werden muss, entscheiden kann, ob er zustimmt oder nicht.

Hinweise und Empfehlungen:

Open Data sowie Daten, die nach IFG angefragt wurden, sollten mit Blick auf ihre potenzielle Nutzung und abhängig vom jeweiligen Material in strukturierter und maschinenlesbarer Form zur Verfügung gestellt werden.

Sie sollten sich finden, durchsuchen, filtern und von anderen Anwendungen weiterverarbeiten lassen. Daten von Regierungsstellen liegen jedoch aktuell oft nur als PDF vor und sind somit nicht ohne Probleme für Forschungszwecke weiter zu verarbeiten. Darüber hinaus sollte die Auffindbarkeit derartiger Daten bspw. durch geeignete Suchmaschinenoptimierung oder Listung an zentralen Stellen (bspw. in Form einer Metadatenbank) verbessert werden.

3 Web Scraping: Datenzugang und Forschungsdatenmanagement

Dem individuellen und institutionellen Zugangsweg zu Big Data ist gemein, dass ihre rechtlichen Grundlagen durch Verträge oder verbindliche Absprachen begründet werden. Für Open Data und die Informationsfreiheit geben gesonderte Gesetze auf Bundes- und Landesebene die rechtliche Rahmung vor.

Hinsichtlich des Web Scrapings bestehen jedoch größere Unsicherheiten, was rechtlich erlaubt und geboten ist. Dies betrifft die Nutzung von Schnittstellen, aber insbesondere auch die Anwendung von Screen Scraping, bei dem für Menschen gemachte Webseiten ausgelesen werden. Daher hat der RatSWD zum gesetzlichen Rahmen dieser Zugangswege ein juristisches Gutachten von der Forschungsstelle RobotRecht der Julius-Maximilians-Universität Würzburg eingeholt. Im Folgenden werden die Potenziale des Web Scrapings für die Forschung skizziert und die zentralen Gesichtspunkte des Gutachtens zum Datenzugang und Datenarchivierung von Big Data zusammengefasst. Das vollständige Gutachten findet sich im Anhang.[16]

→ Gutachten S. 31

3.1 Fallbeispiel: Reputation und Kooperation in anonymen Internetmärkten

Zwischenzeitlich gibt es in den Sozial-, Verhaltens- und Wirtschaftswissenschaften eine Vielzahl von Studien, die sich der Methode des Web Scrapings bedienen, um bspw. eigene Daten mit Internetdaten anzureichern oder diese Internetdaten als eigenständige Datenquelle für deskriptive und analytische, alte und neue Fragestellungen zu nutzen. Stellvertretend für andere Untersuchungen wird im Folgenden eine in ‚American Sociological Review' veröffentlichte Studie von Diekmann et al. (2014) skizziert, die sich mit dem Zusammenhang von Reputation und Kooperation in anonymen Internetmärkten beschäftigt. Interessant ist diese Studie vor allem deshalb, weil seit längerem existierende Fragestellungen nun auf Basis neuer Datenquellen und Messmethoden untersucht werden. Darüber hinaus gibt es eine Vielzahl weiterer Studien, die solche Daten für neue Fragestellungen nutzen (zum Überblick siehe z. B. Gosling und Mason 2015).

Vertrauen, Reziprozität und Reputation gelten als wesentliche Voraussetzungen für die Kooperation in sozialen Austauschbeziehungen, seien sie privater oder geschäftlicher Natur. Ein Austausch unter anonymen Akteuren birgt Risiken, da jeder Akteur mehr oder weniger kooperieren oder sich betrügerisch verhalten kann. Wiederholte Interaktionen derselben Akteure verringern solche Probleme, da die vorliegenden Erfahrungen die gegenseitigen Erwartungen in Hinblick auf zukünftige Interaktionen beeinflussen.

Eine solche Interaktions-Historie ist bei anonymen Internet-Auktionen nicht gegeben, da Verkaufende und Kaufende zumeist nur einmal interagieren. Hinzu kommt, dass sich betrügerisches Verhalten von der einen oder anderen Seite einer strafrechtlichen Verfolgung häufig entzieht. Online-Märkte begegnen diesem Defizit an institutioneller Absicherung von ökonomischen Transaktionen durch die Etablierung von Reputationssystemen. Die Effektivität eines solchen Reputationssystems basiert auf der Bereitschaft von Kaufenden und Verkaufenden, im Anschluss einer Transaktion eine gegenseitige Bewertung vorzunehmen. Dies ist insofern nicht selbstverständlich, da das Reputationssystem als kollektives Gut von allen Akteuren auch dann (kostenfrei) genutzt werden kann, wenn sie selbst nicht kooperieren, d. h. keine Bewertungen abgeben. In ihrer Studie beschäftigen sich Diekmann et al. mit

16 Hinweis: Das Gutachten definiert – abweichend vom vorliegenden Bericht – Screen Scraping an einer Stelle als ‚allgemeiner' als Web Scraping. Der Oberbegriff ‚Web Scraping' steht im Gutachten und im vorliegenden Bericht sowohl für die Nutzung von Schnittstellen als auch für das Auslesen von Webseiteninhalten. Die Aussagen des Gutachtens beziehen sich auf Web Scraping.

den Mechanismen, die eine solche Kooperation begünstigen. Konkret interessieren sie sich erstens dafür, ob Reputation einen Marktwert hat, so dass es für die Akteure auf Internetbörsen rational ist, sich eine gute Reputation aufzubauen. Zweitens beschäftigen sie sich mit dem Bewertungsverhalten der Akteure, da eben diese (realitätsnahen) Bewertungen die Grundlage für ein effektives Reputationssystem und damit die Grundlage für das Funktionieren anonymer Internetmärkte bildet. Als Datenquelle nutzen sie Auktionen auf der deutschsprachigen Version der eBay-Plattform. Die interessierenden Daten wurden mittels Screen Scraping erhoben. Beobachtet wurden Auktionen für zwei ausgewählte Produkte (Mobiltelefone, DVDs). Über einen Beobachtungszeitraum von einem Monat wurden Daten von über einer Million Auktionen für die ausgewählten Produkte gesammelt. Nach Abschluss der Auktionen wurden mittels Screen Scraping folgende Informationen erhoben:

1) Hauptseite der Auktion (item page) inklusive HTML Code,
2) Liste der Bieter und Bieterinnen,
3) Profilseite des Verkäufers/der Verkäuferin und des Käufers/der Käuferin,
4) Angebotsseite des Verkäufers/derVerkäuferin und des Käufers/der Käuferin,
5) frühere Bewertungen der Akteure (Bewertungshistorie).

Aus dieser einmonatigen Vollerhebung wurde inhaltlich begründet eine weitere Auswahl getroffen, so dass für die letztendlichen Analysen 350.000 Auktionen genutzt werden konnten. Die Details der Datenerhebungen wurden in einem Online-Supplement[17] publiziert. Darüber hinaus finden interessierte Forschende die Daten wie auch die Analyseskripte online[18], so dass eine Replikation der Analysen möglich sein sollte.

Der Auktionserfolg wurde über zwei Indikatoren gemessen: (1) Verkauf (ja/nein) und (2) erzielter Verkaufspreis. Die Reputation von Anbietenden und Kaufenden wurden über die Anzahl der positiven und negativen Bewertungen gemessen. Das Bewertungsverhalten nach erfolgreicher Auktion wurde über verschiedene Indikatoren gemessen, wie z.B. Dauer (in Tagen) bis Verkaufende und/oder Kaufende eine Bewertung vornahmen, die Art der Bewertung (positiv, negativ, neutral) und frühere wechselseitige Bewertungen der beiden Akteure. Festzuhalten ist weiterhin, dass zwar das Verhalten von Personen beobachtet wurde, die Beobachtungseinheiten und späteren Analyseeinheiten aber nicht die Personen als solche, sondern die Produkte (bzw. Auktionen) waren.

Als wesentliche Ergebnisse der Studie lässt sich festhalten, dass die These eines Marktwerts der Reputation Unterstützung findet. In Analogie zu anderen Studien finden die Autoren, dass eine positive (negative) Reputation der Verkaufenden einen positiven (negativen) Effekt auf Verkaufswahrscheinlichkeit und Verkaufspreis hat. Weiterhin zeigt sich, dass das Bewertungsverhalten stark durch das Prinzip der Reziprozität geprägt ist: Sobald einer der Akteure eine Bewertung des Gegenübers vorgenommen hat, steigt die Wahrscheinlichkeit, dass der Andere gleichfalls eine Bewertung abgibt, stark an. In Ergänzung zu diesem Reziprozitätsprinzip lässt sich ein von den Autoren als ‚altruistisch' charakterisierter Effekt beobachten. Kaufende neigen bei Verkaufenden mit wenigen positiven Bewertungen eher zu einer positiven Bewertung,[19] während Verkaufende mit sehr vielen positiven Bewertungen eher negativ bewertet werden. Haben sich die Akteure in der Vergangenheit bereits bewertet, verringert sich die Wahrscheinlichkeit einer erneuten Bewertung.

17 https://journals.sagepub.com/doi/suppl/10.1177/0003122413512316 (Zugriff am 15.4.2019).
18 https://www.ethz.ch/content/specialinterest/gess/chair-of-sociology/en/publikationen/data.html
19 Nach Diekmann et al. (2014) könnte dies die Bereitschaft von (zufriedenen) Kaufenden reflektieren, Verkaufende beim Aufbau einer Reputation zu unterstützen.

3.2 Erhebung, Archivierung und Nachnutzung der erhobenen Daten

3.2.1 Datenzugang

Das Gutachten „Web Scraping in der unabhängigen wissenschaftlichen Forschung“ der Forschungsstelle RobotRecht (FoRoRe) diskutiert den Datenzugang und die Datenarchivierung via Web Scraping unter Gesichtspunkten des Wettbewerbsrechts, des Urheberrechts und des allgemeinen Zivilrechts. Dabei ist insbesondere das Urheberrecht (UrhG) für Einschränkungen bei der Nutzung relevant (Vogel und Hilgendorf 2019: 33). Hier adressiert insbesondere die Schrankenregelung des § 60d UrhG Privilegien der nicht-kommerziellen wissenschaftlichen Forschung. Während Web Scraping zu kommerziellen Zwecken zustimmungspflichtig ist und bei Zuwiderhandlung Unterlassungs- und Schadenersatzansprüche drohen, bedarf es für die unabhängige wissenschaftliche Forschung unter bestimmten Voraussetzungen keiner Zustimmung (ebd.: 40 f.).

Das Gutachten arbeitet als Kriterien für die nicht-kommerzielle wissenschaftlich Forschung heraus:

1) Die auszuwertenden Informationen müssen allgemein zugänglich sein. Das trifft auch auf Informationen zu, die erst nach Zahlung eines Entgelts abgerufen werden können. (ebd.: 46)
2) Technische Schutzmaßnahmen, die das Web Scraping verhindern sollen, dürfen nicht überwunden werden. Im Gutachten genannt werden als Schutzmaßnahmen beispielsweise robots.txt-Dateien. (ebd.)
3) Die wissenschaftliche Forschung darf ausschließlich nicht-kommerziellen Zwecken dienen. (ebd.)
4) „Durch den Einsatz von (Web) Scraping-Technologien darf keine technische Schädigung beim Betreiber der Website […] eintreten.“ (ebd.) Eine Schädigung würde zum Beispiel eintreten, „[…] wenn der massenhafte Abruf von Daten, die Serverinfrastruktur stark be- oder sogar überlastet und ein ordnungsgemäßer Betrieb der Website – auch kurzfristig – nicht mehr aufrechterhalten werden kann.“ (ebd.: 44)
5) Der Rechteinhaber hat Anspruch auf Zahlung einer angemessenen Vergütung. Dieser Anspruch kann ausschließlich durch eine Verwertungsgesellschaft – nicht durch den Rechteinhaber selbst – geltend gemacht werden. Schuldner dieser Vergütung ist die Einrichtung, an der die oder der Forschende beschäftigt ist. (ebd.: 45 f.)

Insbesondere der letzte Punkt impliziert, dass vor Durchführen eines Web Scrapings Kontakt mit dem Rechteinhaber der Website aufgenommen werden sollte. Denn in aller Regel werden sich „[…] die Vertragspartner in Vergütungsverhandlungen über Höhe und Art der Vergütung verständigen“ (ebd.: 45).

Die Kontaktaufnahme empfiehlt sich auch deshalb, weil noch nicht abschließend geklärt ist, ob beim grenzüberschreitenden Datentransfer der Handlungsort einer möglichen Urheberrechtsverletzung am Standort des Servers oder am Standort der oder des Scrapenden zu sehen ist und somit womöglich das Urheberrecht am Standort des Servers zu beachten ist (ebd.: 48).

Beim Datenabruf über eine API kommt in der Regel ein Vertrag zwischen Rechteinhabenden und Forschenden zustande. Letztere stimmen den Nutzungsbedingungen zu. Die Rechtmäßigkeit des Web Scrapings richtet sich fortan nach diesem Vertrag (ebd.: 47 f.).

3.2.2 Archivierung und Nachnutzung

Sind im gescrapten Material personenbezogene Daten enthalten, so sind selbstverständlich die datenschutzrechtlichen Bestimmungen der europäischen Datenschutzgrundverordnung (DSGVO) zu beachten (Vogel und Hilgendorf 2019: 47). Personenbezogene Daten sind „[…] alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person […] beziehen" (Art 4 Nr. 1 DSGVO). Unabhängig von der Herkunft der Daten unterliegen Forschende in Deutschland dem deutschen Rechtssystem (Vogel und Hilgendorf 2019: 48). Insbesondere weist das Gutachten darauf hin, dass bei der Pseudonymisierung darauf zu achten ist, dass die Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden können (ebd.: 49). Dies wäre beispielsweise der Fall, wenn Alias-Namen die Klarnamen der dahinterstehenden Personen enthalten.

Zur Archivierung und Nachnutzung der durch Web Scraping gewonnen Daten stellt das Gutachten weiterhin fest, dass nach Abschluss der Forschungsarbeiten eine Löschung (= unwiederbringliche Beseitigung) der Daten beim Forschenden vorgeschrieben ist. Dazu wird Forschenden empfohlen, vorab ein Löschkonzept auszuarbeiten und die Befolgung der Löschpflicht anschließend zu dokumentieren. Der Abschluss der Forschung schließt dabei Handlungen zur Qualitätskontrolle, wie etwa Peer-Review, mit ein (ebd.: 46).

Allerdings ist eine Übermittlung an Bibliotheken, Archive, Museen und Bildungseinrichtungen (privilegierte Institutionen im Sinne des § 60e und § 60f UrhG) zur langfristigen Gewährleistung der Überprüfung der Einhaltung wissenschaftlicher Standards und der Zitier- und Referenzierbarkeit gestattet (ebd.: 46 f). Es existiert noch keine Rechtsprechung, ob Forschungsdatenzentren (FDZ) zu diesen privilegierten Institutionen zählen (ebd.: 41). Das Gutachten kommt zu der Einschätzung, dass FDZ – „[…] soweit sie keine unmittelbaren kommerziellen Zwecke verfolgen – dem Grunde nach als privilegierte Institutionen qualifiziert werden" (ebd.). Aufgrund der Heterogenität der vom RatSWD akkreditierten FDZ können aber keine allgemeinverbindlichen Aussagen getroffen werden (ebd.: 42). Die archivierenden Einrichtungen dürfen das übermittelte Material anderen Forschenden zum Zwecke nicht-kommerzieller Forschung zur Verfügung stellen (ebd.: 46 f.). Unklarheit besteht, ob diese Bereitstellung ausschließlich in einer Form erfolgen darf, die ein Ausdrucken oder Abspeichern ausschließt (ebd.: 41). Nicht zulässig ist in der Regel dagegen die Übermittlung des gescrapten Korpus an wissenschaftliche Zeitschriften (ebd.: 42).

4 Institutionalisierter Datenzugang über (treuhänderische) Drittpartei

Wie in Abschnitt 2 und 3 eruiert, gehen mit einem individuellen Zugang zu privaten Big Data-Quellen zahlreiche Herausforderungen und Unabwägbarkeiten einher. Vorhandene institutionalisierte Datenzugänge und die Datengewinnung über Web Scraping können, auch bei Beachtung der vorangegangenen Empfehlungen, ebenso nicht alle Anforderungen an einen nachhaltigen und nutzungsfreundlichen Datenzugang für die Forschung erfüllen. Der RatSWD empfiehlt daher eine unabhängige Forschungseinrichtung zu etablieren oder eine bestehende mit dem Auftrag zu betrauen, treuhänderisch einen standardisierten Zugang zu anonymisierten (Mikro-) Daten aus öffentlichen und privaten Big Data-Quellen bereitzustellen.

Eine solche Treuhandstelle müsste auf Basis einer vertraglichen Regelung die (Mikro-)Daten aus Big Data-Quellen von bestimmten Unternehmen erhalten. Im Rahmen der Vertragsgestaltung wäre zu regeln, welche Daten zu welchen Konditionen, in welchen Formaten und zu welchen Zeitpunkten übergeben werden. Die Treuhandstelle soll dabei als Broker sowohl die Interessen der Forschenden als auch der Unternehmen repräsentieren. Beispielsweise könnte die Treuhandstelle den Unternehmen den Mehrwert aufzeigen, welcher durch eine Auswertung ihrer Daten durch Forschende entsteht. Hier ist nicht nur ein Reputationsgewinn für die Unternehmen zu nennen; die Auswertung von Unternehmensdaten könnte auch gerade kleineren Unternehmen wichtige Einsichten über ihre Aktivitäten bieten. Die Treuhandstelle könnte Unternehmen weiterhin zur Qualität ihrer Datenquellen oder zu methodischen Fragestellungen beraten.

Internationale Perspektive

In den Niederlanden basiert die Kooperation des Statistikamtes, dem Centraal Bureau voor de Statistiek (CBS), und einzelnen Unternehmen, z. B. aus dem Mobilfunk- oder Energiebereich, auf Kooperationsverträgen, in denen die Mehrwerte für beide Seiten geregelt sind.

Ein anderes Modell ist die Initiative ‚Social Science One', die ihren Ursprung und Fokus im U.S.-amerikanischen akademischen System hat. Eine Kommission etablierter Wissenschaftlerinnen und Wissenschaftler handelt in diesem Modell Datenbereitstellungen mit großen Privatunternehmen unter gemeinsamen Forschungsperspektiven aus und schreibt in der Folge Projekte zur Datennutzung durch Forschende aus. Bislang existiert nur eine Kooperation mit Facebook. Staatliche Stellen sind nicht direkt involviert. Das Projekt geht von der Harvard University aus. Weitere Informationen siehe: https://socialscience.one/overview

Die Treuhandstelle wäre mit der Aufgabe betraut, einen standardisierten Zugang zu anonymisierten (Mikro)-Daten aus Big Data-Quellen für zugelassene Stellen zu entwickeln und umzusetzen. Den Forschenden würde die Treuhandstelle dabei hinsichtlich der zu stellenden rechtlichen Fragen Hilfestellungen geben. Wichtige Charakteristika der Treuhandstelle wären, dass es sich um eine unabhängige und nicht-kommerzielle Einrichtung handelt, die ihre Aufgabe aus einem gemeinnützigen Interesse heraus wahrnimmt. Die erforderliche Infrastruktur soll daher durch die öffentliche Forschungsförderung finanziert werden und könnte beispielsweise ein Thema für die Nationale Forschungsdateninfrastruktur (NFDI)[20] sein.

20 Die Nationale Forschungsdateninfrastruktur (NFDI) ist ein Vorhaben der Gemeinsamen Wissenschaftskonferenz (GWK) von Bund und Ländern. Sie soll die bestehenden Forschungsdateninfrastrukturen in Deutschland miteinander vernetzen und erweitern. Auf diesem Weg sollen nutzungsfreundliche Service-Angebote für die Wissenschaft entstehen. Weitere Informationen finden sich auf: https://www.dfg.de/nfdi

Neben der konkreten Datenbereitstellung hätte die Treuhandstelle u.a. noch folgende Aufgaben:

- Kontakt- und Schnittstelle zwischen Wissenschaft und Produzierenden von Big Data:

  Die Treuhandstelle kann von der Wissenschaft angefragt werden, um bei der Einrichtung eines Zugangs zu konkreten Big Data-Quellen zu unterstützen. Potenziell kann die Treuhandstelle auch eigene Anfragen an Datenproduzierende stellen, um weiteres Big Data-Material für die Forschung zugänglich zu machen.
- Sammlung von bewährten vertraglichen Regelungen zur regelmäßigen Bereitstellung der anfallenden Daten aus Big Data-Quellen
- Klärung der rechtlichen Voraussetzungen, welche Stellen in welchem Rahmen und unter welchen Voraussetzungen zugelassen sind, die Daten aus Big Data-Quellen zu nutzen
- Klärung der technischen Möglichkeiten zur Speicherung, Verarbeitung und Verknüpfung der Daten aus heterogenen Big Data-Quellen
- Erarbeitung von Konzepten zur Anonymisierung und Weitergabe der Daten gemäß EU-Datenschutzgrundverordnung und Bundesdatenschutzgesetz
- Erstellung von geeigneten und standardisierten Metadatenbeschreibungen der verwalteten Daten
- Bereitstellung der Daten in geeigneten und standardisierten Formaten zur nutzungsfreundlichen Weiterverarbeitung an zugelassene Stellen
- Beratung der interessierten Stellen zu konkreten Datenanfragen und Datennutzungen

Bei einer Datenweitergabe über eine Treuhandstelle wäre es grundsätzlich hilfreich, wenn es einen gesetzlichen Rahmen gäbe, der einerseits regelt, unter welchen Bedingungen die Daten an die Treuhandstelle übergeben werden müssen, und der andererseits klärt, welche Voraussetzungen gelten, um die seitens der Treuhandstelle aufbereiteten Daten aus Big Data-Quellen Interessierten zur Verfügung zu stellen.

Dieser gesetzliche Rahmen sollte die Weitergabe der Daten seitens der Treuhandstelle auf bestimmte Nutzendengruppen beschränken, wobei vor allem die Nutzendengruppen Wissenschaft und amtliche Statistik eine besondere Bedeutung erhalten sollen. Der Datenzugang für diese beiden Gruppen sollte jeweils gruppenspezifisch geregelt und gewährt werden:

Nutzendengruppe Wissenschaft:

Die unabhängig arbeitende Wissenschaft hat ein Interesse an der Analyse von Daten aus Big Data-Quellen zur Bearbeitung wissenschaftlicher Fragestellungen und zur Beteiligung an nationalen und internationalen wissenschaftlichen Diskursen. Durch den Zugang zu Daten aus Big Data-Quellen würden sich die Bedingungen der deutschen empirischen Wissenschaft im Hinblick auf ihre internationale Wettbewerbsfähigkeit deutlich verbessern. In diesem Zusammenhang ist es für die Wissenschaft einerseits zentral, dass die Ergebnisse der Analysen publiziert, und andererseits, dass die Analysen von anderen Forschenden mit den gleichen Datensätzen repliziert werden können. Die Datenbereitstellung unter Wahrung dieser Rahmenbedingungen zu ermöglichen, wäre Aufgabe der Treuhandstelle.

Nutzendengruppe amtliche Statistik:

Die amtliche Statistik hat Interesse an der Nutzung von Daten aus Big Data-Quellen, u.a. zur inhaltlichen Ergänzung, qualitativen Verbesserung und zeitlichen Aktualisierung ihrer amtlichen Erhebungen. Potenziell können durch die Nutzung von Big Data auch Erhebungspflichten von Auskunftspflichtigen reduziert werden. Für solche Zwecke wäre sicherzustellen, dass die Treuhandstelle die Mikrodaten in dem erforderlichen Umfang bereitstellt. In den statistischen Ämtern werden diese Daten für die oben angeführten Zwecke verwendet.

Dies kann perspektivisch dazu führen, dass diese Daten in die amtlichen Erhebungen integriert und somit Teil des Erhebungs-, Aufbereitungs- und Veröffentlichungsprogramms werden. Ist dies der Fall, sollte die Möglichkeit bestehen, dass diese um Big Data-Quellen ergänzten Statistiken auch über die Forschungsdatenzentren (FDZ) der statistischen Ämter des Bundes und der Länder zu wissenschaftlichen Zwecke entsprechend der geltenden Regelungen zur Verfügung gestellt werden können. In diesem Sinne hat auch der Statistische Beirat in seinen Empfehlungen zur Fortentwicklung der amtlichen Statistik ausgeführt: „Die Daten der amtlichen Statistik werden heute überwiegend über deren FDZ, unter Beachtung der gesetzlichen Anforderungen, der Wissenschaft zur Verfügung gestellt. Dies muss künftig auch für die von der amtlichen Statistik verarbeiteten neuen digitalen Daten möglich sein. Ferner müssen die Ergebnisse der amtlichen Statistik einer wissenschaftlichen Überprüfbarkeit zugänglich bleiben." (Statistischer Beirat 2018: 11)

Für die jeweils konkrete Ausgestaltung der Datenbereitstellung seitens der Treuhandstelle an die zugelassenen Nutzendengruppen (Wissenschaft und amtliche Statistik) ist es sinnvoll, die derzeitige Infrastruktur und die aktuellen Regelungen der FDZ der Statistischen Ämter des Bundes und der Länder als Vorbild zu nehmen. So könnten für die Datenbereitstellung u.a. die folgenden Bedingungen gelten:

- Der Zugang zu den Daten aus Big Data-Quellen erfolgt für alle Nutzenden der zugelassenen Nutzendengruppen (z.B. wissenschaftlichen Einrichtungen, statistische Ämter) auf Basis eines Antrags, in dem u.a. der Verwendungszweck der Daten benannt werden muss.
- Die Nutzung der Daten unterliegt einer zeitlichen Begrenzung, die je nach Nutzendengruppe unterschiedlich ausfallen kann (für die Wissenschaft z.B. projektgebunden, für die amtliche Statistik z.B. gebunden an die jeweils konkrete Erhebung).
- Die Nutzung der Daten ist für den gleichen Nutzenden wiederholt möglich, sofern der Verwendungszweck zulässig ist.
- Die Nutzung des gleichen (standardisierten) Daten- und Metadatensatzes ist auch für andere Nutzende der zugelassenen Nutzendengruppen (z.B. andere wissenschaftliche Einrichtungen, andere statistische Ämter) möglich.
- Der Zugang zu den Daten ist – entsprechend der gesetzlichen Grundlagen – abhängig von ihrem Anonymisierungsgrad. So wäre z.B. denkbar, dass der Zugang zu aggregierten bzw. faktisch anonymisierten Daten weniger stark reglementiert wird als der Zugang zu gering anonymisierten Mikrodaten.
- Der Zugang zu den Daten richtet sich – entsprechend der gesetzlichen Grundlagen – nach der jeweiligen Nutzendengruppe. So wäre z.B. denkbar, dass die amtliche Statistik manche Daten auf anderen Wegen und in anderen Formaten erhält als die Wissenschaft.
- Publikationen von Ergebnissen, die auf Basis einer Datennutzung entstehen, sind mit entsprechenden Quellenangaben zu versehen.
- Die Möglichkeit, im Rahmen von Datennutzungen mit anderen Einrichtungen der zugelassenen Nutzendengruppen zu kooperieren, ist gegeben.

Hinweise und Empfehlungen

Der RatSWD empfiehlt zur Stärkung der deutschen Forschungsinfrastruktur für die Sozial-, Verhaltens- und Wirtschaftswissenschaften eine unabhängige Forschungseinrichtung zu etablieren oder eine bestehende mit dem Auftrag zu betrauen, treuhänderisch einen standardisierten Zugang zu anonymisierten (Mikro-) Daten aus öffentlichen und privaten Big Data-Quellen bereitzustellen. Die dazu erforderlichen Infrastrukturen könnte ein zentrales Thema für die sich derzeit dynamisch entwickelnde Nationale Forschungsdateninfrastruktur (NFDI) sein.

5 Glossar

AG	Arbeitsgruppe
AGB	Allgemeine Geschäftsbedingungen
AJAX	Asynchronous JavaScript and XML (Konzept der Webprogrammierung)
API	Application Programming Interface (Programmierschnittstelle)
BDSG	Bundesdatenschutzgesetz
BGB	Bürgerliches Gesetzbuch
BGH	Bundesgerichtshof
Crawling	Abruf von Dokumenten im Web, die durch das Folgen von Hyperlinks erreicht werden und deren Hyperlinks ihrerseits als Ausgangspunkt für weitere Abrufe dienen
CSS	Cascading Style Sheets (Computersprache für die Gestaltung digitaler, meist webbasierter Dokumente)
Destatis	Statistisches Bundesamt (Deutschland)
DSGVO	Datenschutz-Grundverordnung (EU)
EuGH	Europäischer Gerichtshof
FDZ	Forschungsdatenzentrum
GWK	Gemeinsame Wissenschaftskonferenz (koordiniert die gemeinsame Wissenschaftsförderung von Bund und Ländern in Deutschland)
HTML	Hypertext Markup Language (textbasierte Auszeichnungssprache)
IFG	Informationsfreiheitsgesetz
IP	Internetprotokoll
LG	Landgericht
LOD	Linked Open Data (im Internet frei und standardisiert abrufbare, eindeutig identifizierbare Daten)
n.F.	neue Fassung
NFDI	Nationale Forschungsdateninfrastruktur (Deutschland)
NSF	National Science Foundation (Forschungsförderung USA)
OLG	Oberlandesgericht
PDF	Portable Document Format (Dateiformat)
RatSWD	Rat für Sozial- und Wirtschaftsdaten
RDF	Ressource Description Framework (formale Sprache zur Bereitstellung von Metadaten im Internet)
RL	Richtlinie
RWI	Leibniz-Institut für Wirtschaftsforschung
Screen Scraping	Gewinnung von Informationen durch gezieltes Extrahieren aus Daten, die zur (typo-)grafischen Darstellung von Inhalten auf dem Endgerät von Benutzerinnen und Benutzern dienen (Unterbegriff von Web Scraping)
Smart Meter	Intelligenter Zähler (digitales Messgerät und Verbrauchszähler für beispielsweise elektrische Energie, Erdgas, Fernwärme oder Wasser)

SMS	Short Message Service (Telekommunikationsdienst für textbasierte Kurznachrichten)
Spidering	Siehe „Crawling“
UrhG	Gesetz über Urheberrecht und verwandte Schutzrechte
UrhWissG	Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft
URI	Uniform Ressource Identifier (Identifikator für Ressourcen im Internet)
URL	Uniform Ressource Locator (Bezeichnungsstandard für Netzwerkressourcen)
UWG	Gesetz gegen den unlauteren Wettbewerb
Web Scraping	Gewinnung von Informationen durch gezieltes Extrahieren aus Daten, die im World Wide Web bereitstehen (siehe auch Unterform ‚Screen Scraping‘ zur Abgrenzung)
XML	Extensible Markup Language (erweiterbare Auszeichnungssprache)

6 Literaturverzeichnis

Adjerid, Idris und Ken Kelley (2018): Big data in psychology: A framework for research advancement. American Psychologist 73(7), 899–917.

An de Meulen, Philipp; Martin Micheli und Sandra Schaffner (2014): Documentation of German Real Estate Market Data – Sample of Real Estate Advertisements on the Internet Platform ImmobilienScout24. RWI Materialien 80. https://www.rwi-essen.de/publikationen/rwi-materialien/327 (Zugriff am 24.05.2019).

Boeing, Geoff und Paul Waddell (2016): New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. https://journals.sagepub.com/doi/full/10.1177/0739456X16664789.

Boelmann, Barbara und Sandra Schaffner (2019): FDZ Data description: Real-Estate Data for Germany (RWI-GEO-RED) – Advertisements on the Internet Platform ImmobilienScout24. RWI Projektberichte. Essen. http://www.rwi-essen.de/media/content/pages/publikationen/rwi-projektberichte/pb_fdz_-_rwi-geo-red_data_description.pdf (Zugriff am 24.05.2019).

Bug, Mathias (2015): Ansätze und Datenquellen in der Kriminalitätsmessung: ein Überblick zu den offen zugänglichen WISIND-Daten. DIW Vierteljahrshefte zur Wirtschaftsforschung 84(2), 69–101. https://doi.org/10.3790/vjh.84.2.5.

Carrière-Swallow, Yan und Felipe Labbé (2013): Nowcasting with Google Trends in an Emerging Market. Journal of Forecasting 32(4), 289–298.

Chen, Le; Alan Mislove und Christo Wilson (2016): An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. https://cbw.sh/static/pdf/amazon-www16.pdf.

Davis, Donald R.; Jonathan I. Dingel; Joan Monras und Eduardo Morales (2019): How Segregated is Urban Consumption? Journal of Political Economy 127(4), 1684-1738. http://faculty.chicagobooth.edu/jonathan.dingel/research/davisdingelmonrasmorales.pdf.

Di Bella, Enrico; Lucia Leporatti und Filomena Maggino (2018): Big Data and Social Indicators: Actual Trends and New Perspectives. Social Indicators Research 135(3), 869–878. https://doi.org/10.1007/s11205-016-1495-y.

Diekmann, Andreas; Ben Jann; Wojtek Przepiorka und Stefan Wehrli (2014): Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets. American Sociological Review 79(1), 65–85.

Edelman, Benjamin (2012): Using Internet Data for Economic Research. The Journal of Economic Perspectives 26(2), 189–206.

Einav, Liran und Jonathan Levin (2014): The Data Revolution and Economic Analysis. in: Lerner, Josh und Scott Stern (Hrsg.): Innovation Policy and the Economy. National Bureau of Economic Research Innovation Policy and the Economy, 1–24.

Fishman, Elliot (2016): Bikeshare: A Review of Recent Literature. Transport Reviews 36(1), 92–113.

Frees, Beate und Wolfgang Koch (2018): ARD/ZDF-Onlinestudie 2018: Zuwachs bei medialer Internetnutzung und Kommunikation. Media Perspektiven 2018(9): 398-413. https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2018/0918_Frees_Koch_2019-01-29.pdf (Zugriff am 12.09.2019).

Gosling, Samuel D. und Winter Mason (2015): Internet research in psychology. Annual Review of Psychology 66, 877–902.

Gyódi, Kristóf (2017): Airbnb and Booking.com: Sharing Economy Competing Against Traditional Firms? Working Paper DELab UW 2017(3). http://www.delab.uw.edu.pl/wp-content/uploads/2017/09/WP_3_2017_K.Gyodi_.pdf.

Hadam, Sandra (2018): Nutzung von Mobilfunkdaten für amtliche Statistiken. Methoden -- Verfahren -- Entwicklungen. Nachrichten aus dem Statistischen Bundesamt 2018(2), 6–9.

Hannak, Aniko; Claudia Wagner; David Garcia; Alan Mislove; Markus Strohmaier und Christo Wilson (2017): Bias in online freelance marketplaces: Evidence from TaskRabbit and Fiverr. CSCW '17 Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM, 1914–1933. https://doi.org/10.1145/2998181.2998327.

Hyunyoung, Choi und Hal Varian (2012): Predicting the Present with Google Trends. Economic Record 88(S1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Kinne, Jan und Janna Axenbeck (2018): Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany, ZEW Discussion Paper No. 18-033. http://ftp.zew.de/pub/zew-docs/dp/dp18033.pdf.

Lazer, David; Ryan Kennedy; Gary King und Allessandro Vespignani (2014): The Parable of Google Flu: Traps in Big Data Analysis. Science 343(6176), 1203–1205.

McCormick, Tyler H.; Hedwig Lee; Nina Cesare; Ali Shojaie und Emma S. Spiro (2017): Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. Sociological Methods & Research 46(3), 390–421. https://doi.org/10.1177/0049124115605339.

McLaren, Nick und Rachana Shanbhogue (2013): Using Internet Search Data as Economic Indicators. Bank of England Quarterly Bulletin 51(2), 134–140.

Moat, Helen Susannah; Chester Curme; Adam Avakian; Dror Y. Kennett; Eugene Stanley und Tobias Preis (2013): Quantifying Wikipedia Usage Patterns Before Stock Market Moves. Scientific Reports 3(1801). https://www.nature.com/articles/srep01801 (Zugriff am 24.05.2019). https://doi.org/10.1038/srep01801.

Morstatter, Fred; Jürgen Pfeffer; Huan Liu und Kathleen M. Carley (2013): Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. Seventh International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPaper/6071 (Zugriff am 24.05.2019).

Ørmen, Jacob (2019): From Consumer Demand to User Engagement: Comparing the Popularity and Virality of Election Coverage on the Internet. The International Journal of Press/Politics 24(1), 49–68. https://doi.org/10.1177/1940161218809160.

Pfeffer, Jürgen; Katja Mayer und Fred Morstatter (2018): Tampering with Twitter's Sample API. EPJ Data Science 7(1), 1–21.

Powell, Ben; Guy Nason; Duncan Elliott; Matthew Mayhew; Jennifer Davies und Joe Winton (2017): Tracking and modelling prices using web‐scraped price microdata: towards automated daily consumer price index forecasting. Journal of the Royal Statistical Society, Series A. Statistics in Society 181(3), 737–756. https://doi.org/10.1111/rssa.12314.

Reinsel, David; John Gantz und John Rydning (2018): Data Age 2025. The Digitization of the World From Edge To Core. IDC White Paper, November 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (Zugriff am 24.05.2019).

Rieckmann, Johannes und Jan-Lucas Schanze (2015): Sicherheitsempfinden in sozialen Medien und Suchmaschinen – ein realistisches Abbild der Kriminalitätsbelastung? DIW Wochenbericht 2015(12), 271–279. https://www.diw.de/documents/publikationen/73/diw_01.c.498951.de/15-12.pdf.

Schmidt, Torsten und Simeon Vosen (2011): Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. Journal of Forecasting 30(6), 565–578.

Schmidt, Torsten und Simeon Vosen (2012): A Monthly Consumption Indicator for Germany Based on Internet Search Query Data. Applied Economics Letters 19(7), 683–687.

Schmidt, Torsten und Simeon Vosen (2013): Forecasting Consumer Purchases Using Google Trends. Foresight: The International Journal of Applied Forecasting (30), 38–41.

Sen, Indira; Fabian Flöck; Katrin Weller; Bernd Weiß und Claudia Wagner (2019): A Total Error Framework for Digital Traces of Humans. Computing Research Repository. Working Paper. July 22, 2019. arXiv:1907.08228.

Slivko, Olga (2018): 'Brain Gain' on Wikipedia: Immigrants Return Knowledge Home. ZEW - Centre for European Economic Research Discussion Paper 18(008). https://doi.org/10.2139/ssrn.3124193.

Statistischer Beirat (2018): Fortentwicklung der amtlichen Statistik. Empfehlungen des Statistischen Beirats für die Jahre 2018 bis 2022. https://www.destatis.de/DE/Ueber-uns/Leitung-Organisation/Statistischer-Beirat/FortentwicklungNov2018_2022_Teil3.pdf (Zugriff am 03.05.2019).

Verma, Amit; Kirill M. Yurov; Peggy L. Lane und Yuliya V. Yurova (2019): An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements, Journal of Education for Business, 94:4, 243–250, https://doi.org/10.1080/08832323.2018.1520685.

Vogel, Paul und Eric Hilgendorf (2019): Web Scraping in der unabhängigen wissenschaftlichen Forschung. Gutachten im Auftrag des Wissenschaftszentrums Berlin für Sozialforschung gGmbH (WZB) – Rat für Sozial- und Wirtschaftsdaten (RatSWD). In: RatSWD: Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement. RatSWD Output 4 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.39. Im Anhang dieser Publikation.

Von Lucke, Jörn und Christian Geiger (2010): Open Government Data. Frei verfügbare Daten des öffentlichen Sektors. Gutachten für die Deutsche Telekom AG zur T-City Friedrichshafen. https://www.zu.de/institute/togi/assets/pdf/TICC-101203-OpenGovernmentData-V1.pdf (Zugriff am 02.05.2019).

Von Schönfeld, Max (2018): Screen Scraping und Informationsfreiheit. Baden-Baden, Nomos.

Yongping, Zhang und Mi Zhifu (2018): Environmental benefits of bike sharing: A big data-based analysis. Applied Energy 220, 296–301. https://doi.org/10.1016/j.apenergy.2018.03.101.

Mitwirkende bei der Erstellung

Mitglieder der AG

Prof. Dr. Thomas K. Bauer *(Co-Vorsitz der AG)*
RWI – Leibniz-Institut für Wirtschaftsforschung, Ruhr-Universität Bochum, RatSWD

Prof. Dr. Michael Eid
Freie Universität Berlin, RatSWD

Hans-Josef Fischer *(Co-Vorsitz der AG)*
Landesbetrieb Information und Technik Nordrhein-Westfalen (IT.NRW), RatSWD

Dr. Fabian Flöck
GESIS – Leibniz-Institut für Sozialwissenschaften

Prof. Dr. Anja Göritz
Albert-Ludwigs-Universität Freiburg, RatSWD

Heike Habla
Statistisches Bundesamt, RatSWD

Prof. Dr. Kai Maaz
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Goethe-Universität Frankfurt am Main, RatSWD

Sabine Ohsmann
Deutsche Rentenversicherung Bund, RatSWD

Prof. Dr. Mark Trappmann
Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit, Otto-Friedrich-Universität Bamberg, RatSWD

Dr. Heike Wirth, GESIS
GESIS – Leibniz-Institut für Sozialwissenschaften, RatSWD

Konsultation

Prof. Dr. Wolfgang Nagel
ScaDS Dresden/Leipzig – Competence Center for Scalable Data Services and Solutions, Technische Universität Dresden

Geschäftsstelle

Dr. Mathias Bug

Dr. Tim Deeken

Dr. Nora Dörrenbächer

Thomas Runge

Anhang

Forschungsstelle RobotRecht
Julius-Maximilians-Universität Würzburg

Web Scraping in der unabhängigen wissenschaftlichen Forschung

Gutachten im Auftrag des Wissenschaftszentrums Berlin für Sozialforschung gGmbH (WZB) – Rat für Sozial- und Wirtschaftsdaten (RatSWD)

Paul Vogel,
Prof. Dr. Dr. Eric Hilgendorf

Würzburg, den 30.08.2018

Zitiervorschlag: Vogel, Paul und Eric Hilgendorf (2019): Web Scraping in der unabhängigen wissenschaftlichen Forschung. Gutachten im Auftrag des Wissenschaftszentrums Berlin für Sozialforschung gGmbH (WZB) – Rat für Sozial- und Wirtschaftsdaten (RatSWD). In: RatSWD: Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement. RatSWD Output 4 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.39.

Inhaltsverzeichnis

Zusammenfassung (Executive Summary) . . . 33

A. Einführung und Problemaufriss . . . 34

B. Rechtliche Vorgaben . . . 36
I. Wettbewerbsrecht . . . 36
II. Urheberrecht . . . 37
1. Datenbankwerkschutz, § 4 Abs. 2 UrhG . . . 37
2. Datenbankherstellerschutz, §§ 87a, 87b UrhG . . . 37
a) Web Scraping: Datenzugang und Forschungsdatenmanagement . . . 37
b) Web Scraping als Verwertungshandlung . . . 38
aa) Rechtsprechung in Deutschland . . . 38
bb) Rechtsprechung des Europäischen Gerichtshofs . . . 39
(1) Entscheidung „Innoweb/Wegener“ . . . 39
(2) Entscheidung „Ryanair“ . . . 40
c) Zwischenfazit . . . 40
3. Schrankenregelung des § 60d UrhG . . . 40
III. Allgemeines Zivilrecht . . . 43
1. Vertragsrecht . . . 43
2. Virtuelles Hausrecht . . . 44

C. Ableitung von Kriterien für das Web Scraping in der Forschung . . . 45
I. Allgemeine Vorgaben für den Einsatz von Web Scraping . . . 45
II. Bedingungen für die Archivierung und langfristige Zurverfügungstellung der gewonnenen Daten . . . 46
1. Urheberrechtliche Anforderungen . . . 46
2. Datenschutzrechtliche Anforderungen . . . 47
III. Exkurs: Praktisches Anwendungsbeispiel . . . 47
1. Bindungswirkung der Nutzungsbedingungen der *Twitter*-API . . . 47
2. Anwendbares Recht bei Auslandsbezug . . . 48
3. Datenschutzrechtliche Implikationen . . . 49
a) Öffentliches Interesse im Sinne des Art. 89 DSGVO . . . 49
b) Personenbeziehbarkeit bei Pseudonymisierung . . . 49

D. Fazit und Ausblick . . . 50

Literaturverzeichnis . . . 51

Zusammenfassung (Executive Summary)

In der sozial-, verhaltens- und wirtschaftswissenschaftlichen Forschung wird vermehrt auf die Technik des Web Scraping zurückgegriffen, um öffentlich im Internet zugängliche Daten automatisiert abzurufen und zusammenzustellen. Das erleichtert das Auffinden von Auffälligkeiten und Korrelationen, was aufgrund der hohen Menge der Daten mit dem bloßen Auge langwierig oder sogar unmöglich wäre.

Allerdings ergeben sich aus juristischer Perspektive einige Bedenken hinsichtlich der Zulässigkeit dieser Verfahrensweise. Vor allem stellen sich Fragen aus dem Wettbewerbs-, Urheber- und allgemeinen Zivilrecht. Das vorliegende Gutachten kommt zu dem Ergebnis, dass das Web Scraping im Kontext der wissenschaftlichen Forschung keine wettbewerbsrechtlichen Implikationen auslöst, da es an der hierzu erforderlichen „geschäftlichen Handlung" in aller Regel fehlen wird. Erhöhten Begründungsbedarf hält die Zulässigkeit aus der Perspektive des Urheberrechts bereit. Bei den auszulesenden Webseiten und Datenbanken handelt es sich meist um geschützte Werke oder Datenbankwerke im Sinne des Urheberrechtsgesetzes (UrhG). Das hat zur Folge, dass jede Nutzung – also auch ein Web Scraping – dem Grunde nach zustimmungs- und vergütungspflichtig ist, bei Zuwiderhandlung drohen empfindliche Unterlassungs- und Schadensersatzansprüche. Der deutsche Gesetzgeber hat nach einigen – sich teils gegenseitig widersprechenden – höchstrichterlichen Entscheidungen des BGH und des EuGH reagiert und für die wissenschaftliche Forschung eine sogenannte Schrankenregelung in das UrhG aufgenommen, die unter bestimmten Voraussetzungen ein Text und Data Mining, wie es beim Web Scraping zur Anwendung kommt, zustimmungsfrei stellt. Für Forschende ist die urheberrechtliche Zulässigkeit daher im Regelfall zu bejahen, wohingegen bei kommerziellen Scrapern noch hohe Rechtsunsicherheit besteht. Vertragliche Regelungen sind bei Scraping-Verfahren häufig nicht existent, auch ein mögliches virtuelles Hausrecht Webseitenbetreibender steht dem Scraping bei vernünftiger Nutzung nicht entgegen.

Aus den Erörterungen zur Rechtslage werden im vorliegenden Gutachten schließlich einige Kriterien für den Einsatz von Web Scraping in der wissenschaftlichen Forschung und sich dem Scraping anschließende Fragen der Archivierung und Zurverfügungstellung abgeleitet.

Im Ergebnis ist festzuhalten, dass Wissenschaftlerinnen und Wissenschaftler sich aus juristischer Perspektive zur Unterstützung ihrer Forschung durchaus dem Verfahren des Web Scraping bedienen können. Dabei müssen sie allerdings einige Voraussetzungen beachten, um nicht mit dem Gesetz in Konflikt zu geraten.

A. Einführung und Problemaufriss

In einer datengetriebenen Gesellschaft, deren weltweites Datenvolumen sich zwischen 2016 und 2025 von 16,1 auf 163 Zettabyte verzehnfachen soll (Seagate n.d.), steigt mit der Menge an Daten und der zunehmenden Möglichkeit der Verknüpfung und Auswertung ihr Wert. Big Data ist längst nicht mehr bloß ein modisches Buzzword, sondern eine etablierte Geschäftspraxis, die allein in Deutschland im Jahr 2018 einen Umsatz von 6,4 Milliarden Euro generieren wird (Bitkom 2018). Aufgrund dieses Aufschwungs, der den häufig zitierten Vergleich von Daten als „neuem Öl“[1] bestätigt, kommt der Erschließung von Datenquellen eine herausragende Bedeutung zu.

Hierzu wurden in den vergangenen Jahren diverse Praktiken entwickelt. Als Web Scraping wird ein technisches Verfahren bezeichnet, das dazu dient, automatisiert Inhalte aus fremden Datenbanken auszulesen. Zu unterscheiden ist das Web Scraping von dem allgemeiner gefassten Begriff des Screen Scaping (dt. „Bildschirm-Auskratzen“). Web Scraping betrifft speziell das Auslesen von Daten aus Webseiten oder auf Webseiten verfügbaren Datenbanken (etwa über Programmierschnittstellen [API]) (von Schönfeld 2018: 25). Die folgenden Ausführungen beziehen sich auf diesen Anwendungsfall. Ein häufiger use case sind Suchmaschinen, die für den Anwender massenhaft Daten von Webseiten – beispielsweise von Hotelanbietern – extrahieren und Kundinnen und Kunden in einer Liste verfügbare Hotelzimmer samt Preis anzeigen. Dieser durch die Technik des Web Scraping entstandene „Anschlussmarkt“ der Vergleichsportale hat eine enorme wirtschaftliche wie gesellschaftliche Relevanz erlangt (Schapiro und Żdanowiecki 2015: 497 (498)). Diese gipfelt darin, dass immer mehr Verbraucherinnen und Verbraucher die Vergleichsportale den eigenen Webseiten der Waren- und Dienstleistungsanbieter vorziehen (Bitkom 2013: 32 f.). Konflikte zwischen den betroffenen Anbietern und Web Scrapern sind daher beinahe vorprogrammiert (Schapiro und Żdanowiecki 2015: 497 (498)).

Doch nicht nur in der Wirtschaft erfreut sich das Web Scraping wachsender Beliebtheit. Auch Disziplinen der wissenschaftlichen Forschung, die auf die Auswertung von großen Datenmengen angewiesen sind, können sich zunehmend die Technik zunutze machen und Inhalte aus öffentlich zugänglichen Datenbanken automatisiert auslesen lassen. Das spart Zeit und ermöglicht eine Fokussierung des Personaleinsatzes auf die anschließende Auswertung der Datenberge. Darüber hinaus erlauben die wachsenden Rechenleistungen das Auffinden von Korrelationen und Auffälligkeiten, die mit dem bloßen Auge oder älteren Techniken aufgrund der großen Menge der Daten übersehen worden wären (vgl. Mörike 2018: 2).

In technischer Hinsicht läuft das Web Scraping in der Regel in zwei wesentlichen Schritten ab: Zunächst wird eine Webseite mittels eines Webbots (auch: Crawler) abgerufen, anschließend werden die dortigen Informationen analysiert und gegebenenfalls extrahiert (von Schönfeld 2018: 51). Dabei stellt sich die Schwierigkeit, dass Internetseiten für ihren Abruf durch Menschen und nicht durch Maschinen konzipiert werden. Häufig liegt daher keine maschinenlesbare Version der abgerufenen Webseite vor, wodurch das Auslesen der Daten erheblich erschwert wird (Kukulenz 2008: 22). Gleichwohl steigen mit der fortschreitenden Entwicklung der Technik die Möglichkeiten, zielgenau relevante von irrelevanten Informationen zu unterscheiden.

1 So in Bezug auf personenbezogene Daten die ehemalige EU-Kommissarin *Meglena Kuneva* in einer Rede vom 31.3.2009 in Brüssel – SPEECH/09/156 („Personal data is the new oil of the Internet and the new currency of the digital world.“).

Zusammenfassend lässt sich Web Scraping als „Oberbegriff für ein Algorithmus-basiertes Verfahren [beschreiben], mit dem im World Wide Web technisch frei zugängliche Informationen und Daten ausgelesen werden. Ein Screen Scraping-Programm simuliert dabei menschliches Nutzungsverhalten, um Zugang zu Webseiten zu erhalten und die dort vorhandenen Informationen und Daten abzurufen und anschließend auszuwerten“ (von Schönfeld 2018: 58).

Fraglich ist jedoch, welche rechtlichen Grenzen dem Web Scraping gesetzt sind und welche Vorgaben – speziell im Kontext der wissenschaftlichen Forschung – zu beachten sind. Hier gestaltet sich die Interessenlage anders als im Fall von kommerziellen Angeboten wie Vergleichsportalen, schließlich erfolgt das Scraping nicht aus rein pekuniären Motiven heraus. Vielmehr ist das Interesse des Datenbank- oder Webseitenanbieters mit dem (Allgemein-) Interesse an einer möglichst ungehinderten wissenschaftlichen Forschung in Einklang zu bringen.

B. Rechtliche Vorgaben

Die rechtliche Zulässigkeit des Web Scraping muss sich im Wesentlichen an drei Rechtsgebieten messen lassen. Erstens könnten sich Probleme auf wettbewerbsrechtlicher Ebene ergeben. Vor allem beim Einsatz des Web Scraping im Bereich von Flugvermittlungen und Preisvergleichsportalen sind Konflikte mit dem Gesetz gegen den unlauteren Wettbewerb (UWG) beinahe vorprogrammiert. Ob sich diese Konflikte auch im Kontext der wissenschaftlichen Forschung ergeben, ist zu Beginn zu erörtern.

Deutlich relevanter und daher entsprechend vertieft zu untersuchen sind zweitens urheberrechtliche Implikationen. Datenbanken und ihre Schöpfenden bzw. Herstellenden genießen nach dem Urheberrechtsgesetz (UrhG) vielfältigen Schutz. Allerdings hat der Gesetzgeber jüngst den Umgang mit fremden Datenbankwerken zu wissenschaftlichen Forschungszwecken privilegiert – nach einer Reihe divergierender höchstrichterlicher Gerichtsentscheidungen könnte die Frage der urheberrechtlichen Zulässigkeit des Web Scraping daher möglicherweise abschließend geklärt sein.

Zuletzt sind auch vertragsrechtliche Erwägungen nicht außen vor zu lassen: Je nach Ausgestaltung der „auszukratzenden" Datenbank steht gegebenenfalls ein Nutzungsvertrag im Raum, der seinerseits die Möglichkeiten des Auslesens beschränken könnte. Namentlich könnte beim unautorisierten Einsatz von Webcrawlern das sogenannte „virtuelle Hausrecht" Webseitenbetreibender verletzt werden.

I. Wettbewerbsrecht

Im Zusammenhang mit der juristischen Bewertung des Web Scraping werden häufig wettbewerbsrechtliche Implikationen erörtert. Auch der BGH hat sich bereits mehrfach mit der wettbewerbsrechtlichen Komponente befasst. Die Rechtsmaterie, die vorwiegend im Gesetz gegen den unlauteren Wettbewerb (UWG) geregelt ist, verbietet bestimmte Handlungen und ahndet einen Verstoß gegen seine Vorschriften mitunter sogar strafrechtlich.

Im Kontext des Web Scrapings werden vor allem die Wettbewerbsverstöße der Nachahmung (§ 4 Nr. 3 UWG), der lauterkeitswidrigen Behinderung (§ 4 Nr. 4 UWG) und der wettbewerbswidrigen Irreführung (§ 5 Abs. 1 S. 2 Nr. 1 UWG) diskutiert. All diese Handlungen sind gemäß § 3 Abs. 1 UWG unzulässig und daher Anknüpfungspunkt für die in den §§ 8 ff. UWG aufgezählten Rechtsfolgen wie Unterlassungs- und Schadensersatzansprüchen. Gemeinsame Voraussetzung all dieser Vorschriften ist allerdings das Vorliegen einer geschäftlichen Handlung im Sinne der Definition des § 2 Abs. 1 Nr. 1 UWG.

> **§ 2 UWG – Definitionen**
>
> (1) Im Sinne dieses Gesetzes bedeutet
>
> 1. „geschäftliche Handlung" jedes Verhalten einer Person zugunsten des eigenen oder eines fremden Unternehmens vor, bei oder nach einem Geschäftsabschluss, das mit der Förderung des Absatzes oder des Bezugs von Waren oder Dienstleistungen oder mit dem Abschluss oder der Durchführung eines Vertrags über Waren oder Dienstleistungen objektiv zusammenhängt; als Waren gelten auch Grundstücke, als Dienstleistungen auch Rechte und Verpflichtungen;

Das Web Scraping müsste demnach einen objektiven Zusammenhang mit der Förderung des Absatzes oder des Bezugs von Waren oder Dienstleistungen oder mit dem Abschluss oder der Durchführung eines Vertrags über Waren oder Dienstleistungen aufweisen. Im wissenschaftlichen Forschungsbetrieb wird ein solcher Bezug in aller Regel nicht vorliegen, sodass wettbewerbsrechtliche Gesichtspunkte dem Web Scraping durch Forschende grundsätzlich nicht entgegenstehen (ebenso Mörike 2018: 4).

II. Urheberrecht

Als eines der zentralen Rechte des geistigen Eigentums wird das Urheberrecht im deutschen Rechtssystem im Wesentlichen durch das Urheberrechtsgesetz (UrhG) kodifiziert. Mit seinen Normen gewährt das Gesetz bei Erfüllung der Schutzvoraussetzungen (insbesondere bei Vorliegen einer persönlichen geistigen Schöpfung) primäre, subjektive Ausschließlichkeitsrechte (sog. Verwertungsrechte) (Rehbinder und Peukert 2018: § 1 Rn. 7, 9). Inhabende von Urheberrechten und verwandten Schutzrechten sollen in ihren persönlichen und geistigen Beziehungen zum Werk geschützt werden und für die Nutzung durch Dritte soll eine angemessene Vergütung gewährleistet werden (Nordemann 2014: Einl. UrhG Rn. 8).

Bei der Technik des Web Scraping könnten entsprechende Schutzrechte der Initiatorin oder des Initiators der auszulesenden Webseite oder Datenbank verletzt werden.

1. Datenbankwerkschutz, § 4 Abs. 2 UrhG

Die Betreibenden der auszulesenden Webpräsenz bzw. Datenbank könnte möglicherweise in den Genuss des Datenbankwerkschutzes nach § 4 Abs. 2 UrhG kommen. Datenbankwerke im Sinne dieser Vorschrift sind Sammelwerke, deren Elemente systematisch oder methodisch angeordnet sind und einzeln zugänglich sind. Voraussetzung für die Qualifikation einer Datenbank oder einer Webseite als „Sammelwerk" ist, dass es sich bei der Auswahl oder Anordnung der Inhalte um eine persönliche geistige Schöpfung handelt (§ 4 Abs. 1 UrhG).[2] Dafür ist vor allem das Vorliegen eines Entscheidungsspielraums maßgeblich, der beispielsweise dann nicht gegeben ist, wenn sich die Auswahl oder Anordnung strikt nach generellen Ordnungskriterien (etwa alphabetisch oder chronologisch) richtet (Marquardt 2014: § 4 Rn. 9; Kotthoff 2013: § 4 Rn. 8). Die reine Erfassung und Aktualisierung von Daten in einer Datenbank inklusive Webpräsenz hat daher nicht die nötige Schöpfungshöhe und weist folglich keinen Werkcharakter auf (Leupold und Demisch 2000: § 4 Rn. 10; Schapiro und Żdanowiecki 2015: 497 (499)). Bei den meisten gescrapten Inhalten handelt es sich mithin nicht um taugliche Datenbankwerke im Sinne der Vorschrift, sodass ein Schutz ihrer Urheberin oder ihres Urhebers nach § 4 Abs. 2 UrhG dem Web Scraping nicht entgegensteht (so auch von Schönfeld 2018: 191).

2. Datenbankherstellerschutz, §§ 87a, 87b UrhG

Die §§ 87a ff. UrhG, die auf der europäischen Datenbankrichtlinie 96/9/EG[3] beruhen, gewähren der oder dem Herstellenden einer Datenbank einen sog. Schutz sui generis für die Investitionen, die zur Herstellung einer Datenbank erforderlich sind (Dreier 2018: vor §§ 87a ff., Rn. 1). Gemäß § 87b Abs. 1 S. 1 UrhG hat die oder der Datenbankherstellende das ausschließliche Recht, die Datenbank insgesamt oder einen wesentlichen Teil von ihr zu vervielfältigen, zu verbreiten und öffentlich wiederzugeben. Dagegen schützen die Vorschriften nicht den Inhalt einer Datenbank selbst, etwa in Gestalt einzelner Daten oder Informationen. Geschützt sind vielmehr deren Zusammenstellung und Systematisierung in Form einer Datenbank (vgl. von Schönfeld 2018: 205). Eine besondere Schöpfungshöhe wird – anders als beim Datenbankwerkschutz nach § 4 Abs. 2 UrhG – nicht vorausgesetzt (Kotthoff 2013: § 87a Rn. 1). Für die Qualifikation als Herstellerin oder Hersteller kommt es nicht darauf an, wer die Datenbank tatsächlich erstellt hat oder pflegt; entscheidend ist vielmehr, wer das wirtschaftliche Risiko – also die Kosten – für ihren Betrieb trägt (Schapiro und Żdanowiecki 2015: 497 (499)).

a) Tauglicher Schutzgegenstand

Um als tauglicher Schutzgegenstand im Sinne des § 87a Abs. 1 UrhG qualifiziert zu werden, muss die Datenbank aus unabhängigen Elementen bestehen, die systematisch oder methodisch angeordnet sind und einzeln mit Hilfe elektronischer Mittel oder auf andere Weise zugänglich sein; zudem muss ihre Beschaffung, Überprüfung oder Darstellung eine wesentliche Investition erfordern.[4]

Im Bereich des E-Commerce stellen Plattformdatenbanken und Datenbanken wie beispielsweise Bewertungsportale, Fahrzeug-Onlinebörsen oder Flugplandatenbanken regelmäßig Datenbanken im

2 Siehe auch *EuGH*, Urteil v. 01.03.2012 – C-604/10 – *Football Dataco*.
3 Richtlinie 96/9/EG des Europäischen Parlaments und des Rates vom 11. März 1996 über den rechtlichen Schutz von Datenbanken, ABl. Nr. L 77, 20 ff.
4 Insoweit ist die Begriffsdefinition deckungsgleich mit derjenigen des Art. 1 Abs. 2 der Datenbankrichtlinie 96/9/EG.

Sinne des § 87a Abs. 1 S. 1 UrhG dar.[5] Ob statischer HTML-Code als Datenbank in diesem Sinne zu qualifizieren ist, wird in der rechtswissenschaftlichen Literatur uneinheitlich bewertet. Trotz des vom Unionsgesetzgeber vorgesehenen weiten Verständnisses des Datenbankbegriffes[6] fällt es schwer, bloßen HTML-Code unter die Voraussetzungen des § 87a Abs. 1 UrhG zu subsumieren. Schließlich dient die Ansammlung digitaler Daten in Form von HTML-Code nicht der Zugänglichmachung einzelner Teile einer Datenbank, sondern allein der Darstellung einer Webseite für die Nutzenden (von Schönfeld 2018: 213; ebenso, Schack 2001: 9 (11 f.)). Gleichwohl können Homepages, die eine Suchfunktion bereitstellen und damit den Nutzenden ermöglichen, gezielt einzelne Elemente der Webseite anzusteuern, die Kriterien einer Datenbank erfüllen (von Schönfeld 2018: 214). In der Regel liegen die Voraussetzungen bei Webseiten, die sich als Sammlung unabhängiger Elemente darstellen (z.B. Online-Lexika oder Online-Enzyklopädien), aber vor (Thum und Hermes 2014: § 87a Rn. 94; Vogel 2017: § 87a Rn. 28). Auch Social-Media-Plattformen wie Facebook und Twitter, denen ebenso wie den meisten anderen dynamischen Webseiten des „Web 2.0" ein Content Management System zugrunde liegt, sind als Datenbanken im Sinne des § 87a Abs. 1 UrhG einzuordnen, da dieses die Inhalte in Datenbanken speichert und über Indizes schnell abrufbar bereithält (Thum und Hermes 2014: § 87a Rn. 95).

Für die Bejahung einer wesentlichen Investition reicht es nach der Rechtsprechung des BGH aus, wenn bei objektiver Betrachtung keine ganz unbedeutenden, von jedermann leicht zu erbringenden Aufwendungen erforderlich waren, um die Datenbank zu erstellen.[7] Je nach benötigtem Aufwand für die Beschaffung, Überprüfung oder Darstellung der Inhalte der auszulesenden Datenbank oder Webseite liegt damit ein tauglicher Schutzgegenstand vor.

b) Web Scraping als Verwertungshandlung

Nach positiver Einstufung einer fremden Inhaltssammlung als Datenbank im Sinne des § 87a Abs. 1 S. 1 UrhG stellt sich im Anschluss die Frage, ob der technische Vorgang des Web Scraping eine die Rechte der oder des Datenbankherstellenden potentiell beeinträchtigende Verwertungshandlung darstellt. Gemäß § 87b Abs. 1 S. 1 UrhG hat die oder der Datenbankherstellende das ausschließliche Recht, die Datenbank oder einen wesentlichen Teil davon zu vervielfältigen, zu verbreiten oder öffentlich zugänglich zu machen. Streitentscheidend ist daher die Frage, ob im Rahmen der automatisierten und massenhaften Abfrage von Daten im Wege des Web Scraping wesentliche Teile einer Datenbank vervielfältigt werden.

aa) Rechtsprechung in Deutschland

Die bisher zum Themenkomplex des Web Scraping ergangene Rechtsprechung lässt weder eine einheitliche Linie erkennen noch ermöglicht sie die Festlegung bestimmter Grundsätze. Das Landgericht Hamburg, das sich im Fall „Automobilbörse-Online" mit der Zulässigkeit eines Scraping-Dienstes befasste, stellte sich auf die Seite des Datenbankherstellers und sah in dem Scraping eine Vervielfältigung eines wesentlichen Bestandteils der Datenbank.[8] Eine solche sei demnach immer dann anzunehmen, wenn durch die Nutzung des Bestandteils „ein erheblicher Schaden für die Amortisation der Investition des Datenbankherstellers" drohe.[9] In der Folgeinstanz hob das Hanseatische Oberlandesgericht die Entscheidung auf[10] und wurde hierin letztlich in der Revisionsinstanz durch den Bundesgerichtshof bestätigt. Der BGH stützte sich dabei auf die Überzeugung, dass der Anbieter der Scraping-Software nicht Haupttäter des Eingriffs in das Datenbankherstellerrechts sei, sondern dieser allein durch die Nutzenden begangen würde.[11]

Entsprechend dieser Argumentation entschied das OLG Hamburg sodann in einem jüngeren Verfahren, in dem sich die Fluggesellschaft *Ryanair* gegen ein Flugvergleichsportal zur Wehr setzte,

5 Vgl. nur *BGH*, Urteil v. 01.12.2010 – I ZR 196/08, Rn. 15 – *Zweite Zahnarztmeinung II* = GRUR 2011, 724; *BGH*, Urteil v. 22.06.2011 – I ZR 159/10, Rn. 27 f. – *Automobil-Onlinebörse* = NJW 2011, 3443.
6 Vgl. *EuGH*, Urteil v. 09.11.2004 – C-444/02, Rn. 22 – *Fixtures-Fußballspielpläne II* = GRUR 2005, 254.
7 *BGH*, Urteil v. 01.12.2010 – I ZR 196/08, Rn. 23 – *Zweite Zahnarztmeinung II* = GRUR 2011, 724 m.w.N.
8 *LG Hamburg*, Urteil v. 09.04.2009 – 310 O 39/08 = BeckRS 2009, 20109.
9 *LG Hamburg*, Urteil v. 09.04.2009 – 310 O 39/08, Rn. 60 (juris) = BeckRS 2009, 20109.
10 *OLG Hamburg*, Urteil v. 18.08.2010 – 5 U 62/09 = GRUR 2011, 728.
11 *BGH*, Urteil v. 22.06.2011 – I ZR 159/10, Rn. 20 ff. – *Automobil-Onlinebörse* = NJW 2011, 3443.

in diesem Punkt zugunsten des Letzteren.[12] Zwar hatte sich der BGH in der Revisionsinstanz auch mit diesem Rechtsstreit zu befassen, allerdings musste er die Frage des Eingriffs in das Datenbankherstellerrecht aufgrund bereits eingetretener Rechtskraft nicht mehr bewerten.[13] Somit galt bis zu diesem Zeitpunkt die Frage der datenbankrechtlichen Zulässigkeit von Web Scraping in Deutschland als geklärt (Schapiro und Żdanowiecki 2015: 497 (499); zustimmend von Schönfeld 2018: 244).

bb) Rechtsprechung des Europäischen Gerichtshofs

Zwei den genannten Urteilen zeitlich nachfolgende Entscheidungen des Europäischen Gerichtshofs (EuGH) stellten die deutsche Rechtsprechung grundlegend in Frage.

(1) Entscheidung „Innoweb/Wegener“

Ausgangspunkt dessen war die Entscheidung „Innoweb/Wegener“, in der der EuGH das Web Scraping unter Verwendung einer sog. „Metasuchmaschine“ untersagte.[14] Ähnlich wie im zuvor ergangenen BGH-Urteil in der Rechtssache „Automobilbörse-Online“ ging es um eine Webseite, auf der Nutzende mit Hilfe einer Suchmaschine anhand verschiedener Kriterien gezielt nach Fahrzeugangeboten suchen konnten. Diese durchsuchte Webseiten Dritter nach den angegebenen Parametern, las deren Datenbanken entsprechend aus und stellte die relevanten Treffer schließlich auf der eigenen Webseite als Suchergebnisse dar (vgl. Schapiro und Żdanowiecki 2015: 497 (499)). Anders als der BGH setzte der EuGH den Schutz des Datenbankherstellers zu einem früheren Zeitpunkt an: Bereits das Verfügbarmachen des Portals sei als Eingriff in das Datenbankherstellerrecht zu qualifizieren, sodass die einzelnen Handlungen der Nutzenden für die Bewertung keine Rolle mehr spielten.[15] Diese unterschiedliche Bewertung ergibt sich vor allem daraus, dass die europäische Datenbankrichtlinie 96/9/EG die dem Datenbankhersteller vorbehaltenen Nutzungshandlungen weiter fasst („Entnahme“ und „Weiterverwendung“, Art. 7 Abs. 2 der RL) als der deutsche Gesetzgeber („Vervielfältigung“, „Verbreitung“ und „öffentliche Wiedergabe“, § 87b Abs. 1 UrhG) (Schapiro und Żdanowiecki 2015: 497 (499)).

Artikel 7 Richtlinie 96/9/EG – Gegenstand des Schutzes

(2) Für die Zwecke dieses Kapitels gelten folgende Begriffsbestimmungen:

a) „Entnahme“ bedeutet die ständige oder vorübergehende Übertragung der Gesamtheit oder eines wesentlichen Teils des Inhalts einer Datenbank auf einen anderen Datenträger, ungeachtet der dafür verwendeten Mittel und der Form der Entnahme;

b) „Weiterverwendung“ bedeutet jede Form öffentlicher Verfügbarmachung der Gesamtheit oder eines wesentlichen Teils des Inhalts der Datenbank durch die Verbreitung von Vervielfältigungsstücken, durch Vermietung, durch Online-Übermittlung oder durch andere Formen der Übermittlung. Mit dem Erstverkauf eines Vervielfältigungsstücks einer Datenbank in der Gemeinschaft durch den Rechtsinhaber oder mit seiner Zustimmung erschöpft sich in der Gemeinschaft das Recht, den Weiterverkauf dieses Vervielfältigungsstücks zu kontrollieren.

Der öffentliche Verleih ist keine Entnahme oder Weiterverwendung.

Gerade das Auffangtatbestandsmerkmal „durch andere Formen der Übermittlung“ im Sinne von Art. 7 Abs. 2 lit. b der Richtlinie wird traditionell weit verstanden und wurde vom EuGH auch in der vorliegenden Entscheidung angewandt (vgl. Schapiro und Żdanowiecki 2015: 497 (499) m.w.N.). Dabei legt der EuGH anders als zuvor der BGH weniger Wert auf die technische Funktionsweise, sondern misst vor allem einem möglichst hohen Investitionsschutz zugunsten der oder des Datenbankherstellenden Relevanz bei (Schapiro und Żdanowiecki 2015: 497 (500)).[16] Im konkreten Fall bewertete der EuGH das Web Scraping durch den beklagten Plattformbetreiber folglich als Eingriff in das Datenbankherstellerrecht des Klägers, weshalb dieser Unterlassungs- und Schadensersatzansprüche geltend machen konnte (vgl. nach deutschem Recht § 97 UrhG).

12 *OLG Hamburg*, Urteil v. 24.10.2012 – 5 U 38/10, Rn. 183 f. (juris) = GRUR-RS 2012, 22946.
13 *BGH*, Urteil v. 30.4.2014 – I ZR 224/12, Rn. 20 – *Flugvermittlung im Internet* = MMR 2014, 740.
14 *EuGH*, Urteil v. 19.12.2013 – C-202/12 – *Innoweb/Wegener* = MMR 2014, 185.
15 *EuGH*, Urteil v. 19.12.2013 – C-202/12, Rn. 37 ff. – *Innoweb/Wegener* = MMR 2014, 185.
16 *EuGH*, Urteil v. 19.12.2013 – C-202/12, Rn. 36 – *Innoweb/Wegener* = MMR 2014, 185.

Gleichwohl eignet sich das „Innoweb/Wegener"-Urteil nicht als Präzedenzfall mit grundsätzlicher Bedeutung für alle Scraping-Situationen: Der EuGH selbst betonte in seiner Entscheidung stets die Besonderheiten des konkreten Einzelfalls und bezeichnete die streitgegenständliche Plattform immer als „spezialisierte Metasuchmaschine" (von Schönfeld 2018: 246). Darüber hinaus stellte er Kriterien zur Reichweite der Entscheidung auf: Die Ausführungen gelten nur für eine Metasuchmaschine, „die dem Endnutzer ein Suchformular zur Verfügung stellt, das im Wesentlichen dieselben Optionen wie das Suchformular der Datenbank bietet", „die Suchanfragen der Endnutzer, in Echtzeit' in die Suchmaschine übersetzt", und „dem Endnutzer die gefundenen Ergebnisse unter dem Erscheinungsbild ihrer Website präsentiert".[17]

(2) Entscheidung „Ryanair"

Etwa ein Jahr später gab eine neuerliche Entscheidung des EuGHs weiteren Zündstoff in die Diskussion um die Zulässigkeit des Web Scraping. Nachdem *Ryanair* auch in den Niederlanden gegen ein Flugvergleichsportal vorgegangen ist, gab der EuGH – anders als zuvor noch der BGH – der Klägerin Recht.[18] Die Begründung hierfür unterschied sich allerdings wesentlich von derjenigen im Fall „Innoweb/Wegener". Die niederländischen Instanzgerichte hatten nämlich entschieden, dass die Flugbuchungsdatenbank auf der Webseite von *Ryanair* aufgrund fehlender wesentlicher Investitionen in die Erfassung und Aufzeichnung vorhandener Daten in eine Datenbank kein Datenbankwerk und keine Datenbank im urheberrechtlichen Sinne darstellt[19] – an diese Feststellungen war der EuGH bei seiner Entscheidung gebunden (Schapiro und Żdanowiecki 2015: 497 (500)).

Das Gericht musste sich sodann mit der ihm vorgelegten Frage befassen, ob eine Datenbank, die wie die von *Ryanair* keinen Schutz im Sinne der Datenbankrichtlinie genießt, dennoch rechtmäßigerweise im Einklang mit den Schrankenregelungen der Art. 6 und 8 der Richtlinie ausgelesen werden darf. Das war deshalb von Bedeutung, da die AGB von *Ryanair*, denen die Nutzenden vor Absendung seiner Suchanfrage zustimmen müssen, das Screen Scraping ausdrücklich untersagen – eine solche Untersagung beeinträchtigt aber das Recht des Nutzenden nach Art. 6 Abs. 1 der Richtlinie und wäre daher gemäß Art. 15 der Richtlinie unwirksam. Der EuGH entschied zugunsten der klagenden *Ryanair Ltd.*, dass deren Datenbank in diesem Kontext insgesamt nicht dem Anwendungsbereich der Richtlinie (und damit auch nicht dem Verbot des Art. 15) unterfällt und die Klägerin ihren Nutzenden daher ein Scraping vertraglich untersagen durfte.[20] Insofern lag zwar kein Verstoß gegen ein Urheber- oder Leistungsschutzrecht seitens des Vergleichsportals vor, dennoch konnte es von *Ryanair* wirksam aufgrund des vertraglichen Verbots auf Unterlassung in Anspruch genommen werden (Schapiro und Żdanowiecki 2015: 497 (500)).

c) Zwischenfazit

Nachdem die Frage der urheberrechtlichen Zulässigkeit des Web Scraping in Deutschland nach der BGH-Rechtsprechung grundsätzlich mit „ja" beantwortet werden konnte, brachte der EuGH mit seinen beiden Entscheidungen diese Rechtsauffassung gründlich ins Wanken. Das oberste Gericht der Europäischen Union maß dem Investitionsschutz des Datenbankherstellers weitaus höhere Bedeutung zu, als dies noch die deutschen Gerichte getan haben. Zumindest in Bezug auf kommerzielle Scraping-Dienstleistungen bleibt daher eine Reaktion des BGH auf die jüngste EuGH-Rechtsprechung mit Spannung abzuwarten.

3. Schrankenregelung des § 60d UrhG

Im hier relevanten Kontext der wissenschaftlichen Forschung kam der deutsche Gesetzgeber dem BGH aber zuvor. Das deutsche Urheberrecht erklärt in den §§ 44a ff. UrhG zahlreiche – an sich urheberrechtsrelevante – Nutzungen für zulässig (Rehbinder und Peukert 2018: § 1 Rn. 17). Ein Urheberrechtsverstoß mit den Konsequenzen der §§ 97 ff. UrhG (Anspruch der Rechteinhaber auf Unterlassung, Schadensersatz etc.) liegt bei Eingreifen einer solchen sogenannten „Schranke" nicht

17 *EuGH*, Urteil v. 19.12.2013 – C-202/12, Ls. – *Innoweb/Wegener* = MMR 2014, 185.
18 *EuGH*, Urteil v. 15.01.2015 – C-30/14 – *Ryanair* = MMR 2015, 189.
19 *Gerechtshof Amsterdam*, Urteil v. 13.03.2012 – 200.078.395, 4.13 = ECLI:NL:GHAMS:2012:BW0096.
20 *EuGH*, Urteil v. 15.01.2015 – C-30/14, Rn. 39 – *Ryanair* = MMR 2015, 189.

vor. Mit dem Urheberrechts-Wissensgesellschafts-Gesetz (UrhWissG)[21], das am 01.03.2018 in Kraft getreten ist, hat der Gesetzgeber neue Privilegien für Nutzungen urheberrechtlich geschützter Werke für Unterricht, Wissenschaft und Institutionen aufgenommen. Dazu gehört auch die neu geschaffene Schranke des § 60d UrhG, der das Text und Data Mining privilegiert.

Anknüpfungspunkt und Gegenstand der automatisierten Auswertung ist ausweislich des § 60d Abs. 1 UrhG eine Vielzahl von Werken. Dieser Begriff, den das Gesetz als „Ursprungsmaterial" legaldefiniert, meint eine große und zunächst unsortierte Text- und Datenmenge von geschützten Inhalten aller Art (Hagemeier 2018: § 60d Rn. 6). Dieses Ursprungsmaterial wird sodann maschinenlesbar gemacht, indem es beispielsweise normalisiert, strukturiert und ggf. umgewandelt wird (etwa von PDF-Dokumenten in XML-Datensätze).[22] Das daraus entstehende „Korpus" ist sodann die auszuwertende Datensammlung, die automatisiert beispielsweise auf statistische Häufigkeiten oder Korrelationen untersucht wird.[23]

Grundvoraussetzung für die Anwendbarkeit der Schrankenregelung ist, dass die wissenschaftliche Forschung ausschließlich nicht-kommerziellen Zwecken dient (§ 60d Abs. 1 S. 2 UrhG). Das ist dann der Fall, wenn die Forschung „nicht gewinnorientiert" oder „in staatlich anerkanntem Auftrag im öffentlichen Interesse" erfolgt (Hagemeier 2018: § 60d Rn. 8). Forschung, die ein Unternehmen betreibt, um Waren oder Dienstleistungen zu entwickeln und diese dann zu vermarkten, dient dagegen kommerziellen Zwecken.[24] Das erstellte Korpus (aber nicht das Ursprungsmaterial (Hagemeier 2018: § 60d Rn. 16)) darf weiterhin gemäß § 60d Abs. 1 S. 1 Nr. 2 UrhG einem abgegrenzten Kreis von Personen für die gemeinsame wissenschaftliche Forschung sowie einzelnen Dritten zur Überprüfung der Qualität wissenschaftlicher Forschung (etwa über ein gemeinsames Intranet (Mörike 2018: 3)) zugänglich gemacht werden. Ob darüberhinausgehende Formen der Vervielfältigungen zulässig sind, geht aus dem Gesetzeswortlaut nicht hervor. Bei strenger Betrachtung des Wortlauts würde man unter einem Zugänglichmachen lediglich das Bereithalten zur Ansicht verstehen, sodass das Korpus nur in Datenformaten bereitgestellt werden dürfte, die ein Ausdrucken oder Abspeichern ausschließen (Hagemeier 2018: § 60d Rn. 15). Das erneute Abspeichern, Ausdrucken oder Versenden per E-Mail (Leupold und Demisch 2000: 379 (385)) wäre nämlich eine Vervielfältigung im Sinne des § 16 UrhG, die vom Wortlaut des § 60d Abs. 1 S. 1 Nr. 2 UrhG gerade nicht erfasst ist (Hagemeier 2018: § 60d Rn. 15). Da diese Problemstellung an einen Rechtsstreit über die Zugänglichmachung von Werken über elektronische Leseplätze in Bibliotheken erinnert, in dem der EuGH zwar feststellte, dass eine Wiedergabe im Sinne der zugrundeliegenden Richtlinie kein Ausdrucken oder Abspeichern auf einem Datenträger erlaubt, Anschlussnutzungen aber über Schrankenregelungen erlaubt sein könnten,[25] sieht ein Autor in der rechtswissenschaftlichen Fachliteratur die Schrankenprivilegierung des § 60c Abs. 1 und Abs. 3 UrhG als Beleg dafür, dass das Korpus in einem speicherfähigen und ausdruckbaren Datenformat zur Verfügung gestellt werden darf (Hagemeier 2018: § 60d Rn. 15). Da der Gesetzeswortlaut insoweit aber ungenau ist, sollte diese Interpretation bis zu einer gesetzgeberischen oder gerichtlichen Klärung sicherheitshalber mit Vorsicht genossen werden.

Zuletzt müssen das Korpus und etwaige Vervielfältigungen des gesammelten Materials gemäß § 60d Abs. 3 UrhG nach Abschluss der Forschungsarbeiten gelöscht werden. Eine Übermittlung an eine Bibliothek oder vergleichbare privilegierte Institution im Sinne der §§ 60e und 60f UrhG ist aber zulässig. Der Wortlaut der Normen umfasst als privilegierte Institutionen neben Bibliotheken nur Archive, Museen und Bildungseinrichtungen. Ob vom RatSWD akkreditierte Forschungsdatenzentren (FDZ) als privilegierte Institutionen im Sinne des Gesetzes qualifiziert werden können, ist durch Auslegung zu ermitteln. Unter „Bibliotheken" im Sinne des § 60e Abs. 1 UrhG versteht das Gesetz „öffentlich zugängliche Bibliotheken, die keine unmittelbaren oder mittelbaren kommerziellen Zwecke verfolgen".

21 Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft vom 01.09.2017, BGBl. I 2017, S. 3346.

22 Amtliche Gesetzesbegründung, BT-Drs. 18/12329, S. 40.

23 BT-Drs. 18/12329, S. 40.

24 BT-Drs. 18/12329, S. 39.

25 Vgl. *EuGH*, Urteil v. 11.09.2014 – C-117/13 = MMR 2014, 822; vorgehend *BGH*, Beschluss v. 20.09.2012 – I ZR 69/11 – Elektronische Leseplätze I = MMR 2013, 529; nachgehend *BGH*, Urteil v. 16.04.2015 – I ZR 69/11 – Elektronische Leseplätze II = MMR 2015, 820.

Die öffentliche Zugänglichkeit ist im Sinne der Norm des § 15 Abs. 3 UrhG zu verstehen (Dreier 2018: § 61 Rn. 16). Demnach muss die Einrichtung primär im Dienste der Gesellschaft handeln, was nicht dadurch ausgeschlossen ist, dass ihre Benutzung an bestimmte Voraussetzungen gebunden ist (z.B. eine Mitgliedschaft) (Hagemeier 2018: § 60e Rn. 19). Dagegen darf der Nutzendenkreis nicht auf einen bestimmten Personenkreis beschränkt werden – eine Reduzierung der Zugänglichkeit auf einen bestimmbaren Personenkreis (etwa die Studierenden einer Universität) ist aber zulässig (ebd.). FDZ sind generell für die unabhängige wissenschaftliche Forschung zugänglich. Es erfolgt daher keine Beschränkung auf einen bestimmten Personenkreis, sondern vielmehr auf einen Zweck der Nutzung. Insofern sind FDZ als öffentlich zugänglich zu qualifizieren. Ob man in ihnen Bibliotheken im Sinne des § 60e UrhG oder Archive gemäß § 60f UrhG sieht, ist für das Ergebnis letztlich unerheblich, da § 60d Abs. 3 S. 2 UrhG insoweit keine Unterscheidung trifft. FDZ können daher – soweit sie keine unmittelbaren oder mittelbaren kommerziellen Zwecke verfolgen – dem Grunde nach als privilegierte Institutionen qualifiziert werden, wobei sich die Beurteilung für einzelne FDZ nach ihrer konkreten Ausgestaltung richtet; allgemeinverbindliche Aussagen können aufgrund der Heterogenität der vom RatSWD akkreditierten Stellen nicht getroffen werden. Zu beachten ist überdies, dass aufgrund des jungen Alters der Vorschriften (Inkrafttreten: 01.03.2018) noch keine konkretisierende Rechtsprechung existiert und die hier vorgenommene Auslegung nur eine Einschätzung der Gutachter darstellt.

Auch eine Übermittlung des mit Hilfe von Techniken des Text und Data Mining erstellten Korpus an publizierende Journals müsste sich an der Vorschrift des § 60d Abs. 3 S. 2 UrhG messen lassen und wird nach den o.g. Kriterien nicht zulässig sein, da Journals bzw. deren Herausgeber keine öffentlich zugänglichen Institutionen im oben erläuterten Sinne darstellen.

Mit der Erlaubnis zur Übermittlung an privilegierte Institutionen will die Vorschrift einen Ausgleich zwischen dem Interesse der Forschenden und dem Interesse der Urheber und Verleger herstellen, immerhin müssen die Forschenden die für ihre Forschung verwendeten Inhalte weiterhin verfügbar halten, um etwa die Überprüfung der Einhaltung wissenschaftlicher Standards zu ermöglichen; auf der anderen Seite haben aber vor allem Wissenschaftsverlage ein Interesse daran, dass keine parallelen Artikeldatenbanken entstehen.[26] Allerdings dürfen ausweislich der Gesetzesbegründung Forschende selbst den Korpus und das Ursprungsmaterial ausdrücklich nicht mehr aufbewahren.[27] Dabei adressiert das Gesetz die Löschpflicht an den von der Schrankenregelung nach § 60d Abs. 1 UrhG Privilegierten. Bei mehreren Forschenden (etwa in einem Konsortium) ist davon auszugehen, dass der Konsortialführer die Löschung bei allen Mitforschenden zu überwachen hat. Hierzu enthält allerdings weder das Gesetz noch die Gesetzesbegründung konkrete Hinweise.

Unberührt bleibt von der neuen Schrankenregelung das Verbot der Umgehung technischer Schutzmaßnahmen im Sinne des § 95a UrhG. Sichert der Betreiber einer Webseite seine Inhalte beispielsweise durch eine *robots.txt*-Datei[28] gegen Web Scraping ab, darf der Scraper diese Schutzmaßnahme nicht umgehen (Mörike 2018: 3). Ein eigentlich bestehender Anspruch des Nutzenden gegen den Rechteinhaber gemäß § 95b Abs. 1 S. 1 Nr. 11 UrhG auf Ermöglichung der Durchsetzung seiner Rechte aus § 60d UrhG gilt gemäß § 95b Abs. 3 UrhG nur für offline zugänglich gemachte Werke und hat für Zwecke des Web Scraping daher keine Relevanz (Raue 2017: 656 (658)).

Zusammenfassend lässt sich festhalten, dass im Kontext der wissenschaftlichen Forschung die Technik des Web Scraping trotz der Rechtsprechung des EuGH aus urheberrechtlicher Sicht als zulässig zu bewerten ist, sofern die Voraussetzungen des § 60d UrhG erfüllt sind und dessen Bedingungen eingehalten werden (so auch Mörike 2018: 2 f). Zu beachten ist überdies, dass die §§ 60a bis 60h UrhG zunächst nur bis zum 28.02.2023 gültig sind und gemäß §142 Abs. 2 UrhG nach diesem Tag nicht mehr anzuwenden sind. Dieser Umstand ist damit zu begründen, dass 2022 eine Evaluierung der Gesetzesänderungen vorgesehen ist und im Anschluss über die weitere Gültigkeit oder ggf. erforderliche Modifizierungen entschieden werden soll.[29]

26 BT-Drs. 18/12329, S. 41.
27 BT-Drs. 18/12329, S. 41 f.
28 Vgl. zu technischen Schutzmöglichkeiten gegen Web Scraping von Schönfeld 2018: 60 ff.
29 BT-Drs. 18/12329, S. 49.

III. Allgemeines Zivilrecht

Neben wettbewerbs- und urheberrechtlichen Implikationen werden häufig auch Fragen des allgemeinen Zivilrechts, insbesondere unter dem Stichwort des „virtuellen Hausrechts", diskutiert.

1. Vertragsrecht

Betreiber von Webseiten können durch einen Vertrag Bedingungen für die Benutzung ihres Internetauftritts aufstellen und damit beispielsweise das automatisierte Auslesen ihrer Datenbanken im Wege des Web Scraping untersagen.[30] Voraussetzung für eine entsprechende Verpflichtung des Nutzenden, das Scraping zu unterlassen, ist aber ein wirksamer Vertragsschluss. Ein solcher kommt beispielsweise dann zustande, wenn die Webseite nur nach vorheriger Registrierung mittels Logins durch eine Nutzendenkennung aufgerufen werden kann und im Rahmen der Registrierung den Nutzungsbedingungen zugestimmt werden muss. Hingegen reicht der bloße Aufruf einer Webseite, die lediglich einen einseitigen Hinweis auf die Nutzungsbedingungen des Anbieters enthält, für einen wirksamen Vertragsschluss nicht aus (Deutsch 2009: 1027 (1028)).[31]

Sollte ersterer Fall gegeben sein und ein wirksamer Nutzungsvertrag zwischen dem Anbieter der Webseite und ihren Nutzenden zustande gekommen sein, richtet sich die Rechtmäßigkeit der das Scraping untersagenden Klausel nach dem in den §§ 305 ff. BGB geregelten Recht der Allgemeinen Geschäftsbedingungen (AGB). Danach ist erste Voraussetzung für die Wirksamkeit der Bestimmung, dass es sich bei ihr um eine AGB im Sinne des § 305 Abs. 1 BGB handelt, wenn also Vertragsbedingungen für eine unbestimmte Vielzahl von Verträgen vorformuliert und einseitig von einer Vertragspartei gestellt wurden, ohne im Einzelnen ausgehandelt worden zu sein. Das ist bei Nutzungsbedingungen auf Webseiten in aller Regel der Fall.

Weiterhin muss die Klausel wirksam in den Vertrag einbezogen worden sein. Dafür muss den Nutzenden der Webseite die Kenntnisnahme der Klausel in zumutbarer Weise möglich sein (§ 305 Abs. 2 BGB). Versteckte Links zu den Nutzungsbedingungen, die für den Besucher der Webseite nicht ohne weiteres erkennbar sind, erfüllen diese Voraussetzung nicht, sodass die Bedingungen nicht Vertragsbestandteil werden.

In einem dritten Schritt wird die Klausel nach den Normen der §§ 307 ff. BGB auf ihre inhaltliche Wirksamkeit hin überprüft. Eine Vertragsbestimmung ist demnach unwirksam, wenn sie gegen gesetzliche Bestimmungen verstößt und/oder den Vertragspartner unangemessen benachteiligt. Für die Beurteilung ihrer inhaltlichen Wirksamkeit kommt es auf die konkrete Formulierung der Klausel an, sodass hierüber keine allgemeingültigen Aussagen getroffen werden können (ebenso Mörike 2018: 4).

Speziell im Kontext des Web Scraping in der wissenschaftlichen Forschung ist allerdings noch eine weitere Voraussetzung zu beachten: Neben dem § 60d UrhG (dazu oben) wurde mit dem UrhWissG auch ein § 60g UrhG eingefügt, dessen Abs. 1 regelt, dass vertragliche Bestimmungen die in § 60d UrhG erlaubten Nutzungen nicht beschränken dürfen. Der Rechteinhaber könnte sich folglich auf eine Klausel, die Web Scraping generell untersagt, nicht berufen, wenn dieses unter den Voraussetzungen des § 60d UrhG erfolgt (Mörike 2018: 4). Er ist allerdings nicht gehindert, technische Schutzmaßnahmen zu ergreifen, die ein Web Scraping erschweren oder verhindern. Bei online bereitgestellten Werken gilt gemäß § 95b Abs. 3 UrhG die Pflicht des § 95b Abs. 1 S. 1 Nr. 11 UrhG, dem Text oder Data Miner die Nutzung zu den in § 60d UrhG privilegierten Zwecken zu ermöglichen, nicht.

30 So etwa in den Nutzungsbedingungen von *Ryanair*, abrufbar unter https://www.ryanair.com/de/de/CorporateLinks/nutzungsbedingungen, Punkt 3 (abgerufen am 20.08.2018).

31 *OLG Frankfurt/M.*, Urteil v. 05.03.2009 – 6 U 221/08.

2. Virtuelles Hausrecht

Eine weitere zivilrechtliche Hürde könnte das sogenannte virtuelle Hausrecht sein, dessen Existenz und Konturen im Wesentlichen seit zwei Entscheidungen des LG Bonn[32] und des OLG Köln[33] in den Jahren 1999 und 2000 sowie des OLG München[34] im Jahr 2007 diskutiert werden (allgemein zur Diskussion vgl. Maume 2007: 620 (623 ff.)). Das herkömmliche, „analoge" Hausrecht wird allgemein aus den Eigentums- und Besitzregelungen der §§ 903, 1004 BGB bzw. §§ 858 ff. BGB hergeleitet.[35] Das virtuelle Hausrecht soll beispielsweise dem Betreiber einer Webseite erlauben, Nutzende durch technische Maßnahmen vom Aufruf oder der Benutzung seines Internetauftritts auszuschließen.[36] Voraussetzung dafür ist aber das Vorliegen eines sachlichen Grundes, willkürliche Verbannungen bestimmter IP-Adressen sind unzulässig (Mörike 2018: 4).[37]

Ein solcher Verstoß gegen das Willkürverbot könnte auch dann vorliegen, wenn sich der ausgeschlossene Scraper wie ein typischer menschliche Nutzende verhält, weil er beispielsweise nur eine einzelne Abfrage in der konkreten Datenbank vornimmt (von Schönfeld 2018: 341). Unabhängig davon, ob man ein virtuelles Hausrecht in direkter oder analoger Anwendung der genannten Vorschriften anerkennt, stellt sich die Frage, inwieweit potentielles Web Scraping damit verhindert werden kann. Das virtuelle Hausrecht wurde entwickelt, um Störer des normalen Betriebsablaufs (etwa durch wiederholte Beleidigungen in einem Online-Forum) von der weiteren Benutzung auszuschließen. Sinn und Zweck des virtuellen Hausrechts ist mithin die Unterbindung weiterer Störungen. Maßgeblich ist also die Frage, ob die Technik des Web Scraping ein „normales" Verhalten oder eine Störung des Betriebsablaufs darstellt (überzeugend ebd.: 341 f.).

Bei Nutzung einer Schnittstelle zur Anwendungsprogrammierung (API) zum Abruf von Daten kann bei Einhaltung der dafür gestellten Bedingungen von einem „störenden" Einwirken auf den Betriebsablauf keine Rede sein, schließlich hat der Dienstanbieter sie zu diesem Zweck ja in der Regel bereitgestellt. Doch auch wenn kein Zugriff auf eine API besteht, wird man die automatisierte Datensammlung grundsätzlich nicht als unredliches oder unnormales Nutzendenverhalten qualifizieren können; eine solche Betrachtung würde vor allem Start-ups im Bereich von innovativen Informationsdiensten unangemessen beeinträchtigen und damit Innovation hemmen (ebd.: 342). Eine Grenze wird aber dann zu ziehen sein, wenn der massenhafte Abruf von Daten die Serverinfrastruktur zu stark be- oder sogar überlastet und ein ordnungsgemäßer Betrieb der Webseite oder Datenbank – auch kurzzeitig – nicht mehr aufrechterhalten werden kann (ebd.: 343).

Übereinstimmend mit Meinungen aus der rechtswissenschaftlichen Literatur kann daher davon ausgegangen werden, dass der Figur des virtuellen Hausrechts im Kontext von Web Scraping keine besonders hohe Bedeutung beizumessen ist und diesem bei rationalem Einsatz nicht entgegenstehen wird (ebenso Mörike 2018: 4; von Schönfeld 2018: 343).

32 *LG Bonn*, Urteil vom 16. 11. 1999 – 10 O 457/99 = MMR 2000, 109.
33 *OLG Köln*, Beschluss v. 25.08.2000 – 19 U 2/00 = MMR 2001, 52.
34 *OLG München*, Urteil vom 26.6.2007 – 18 U 2067/07 = MMR 2007, 659.
35 Vgl. statt vieler *BGH*, Urteil v. 08.11.2005 – KZR 37/03, Rn. 23 ff. – *Hörfunkrechte* = NJW 2006, 377.
36 *OLG Köln*, Beschluss v. 25.08.2000 – 19 U 2/00 = MMR 2001, 52; gegen das Bedürfnis eines virtuellen Hausrechts Redeker 2007: 265 (266).
37 Vgl. *OLG Köln*, Beschluss v. 25.08.2000 – 19 U 2/00 = MMR 2001, 52.

C Ableitung von Kriterien für das Web Scraping in der Forschung

Im Folgenden sollen aus der oben erfolgten Darstellung der Rechtslage einige Kriterien für den Einsatz von Web-Scraping-Technologien im Rahmen der unabhängigen wissenschaftlichen Forschung abgeleitet werden.

I. Allgemeine Vorgaben für den Einsatz von Web Scraping

Als zentrales Kriterium für den Einsatz von Scraping-Verfahren aus juristischer Perspektive kann festgehalten werden, dass die auszuwertenden Informationen allgemein zugänglich sein müssen. Die Überwindung von technischen Schutzmaßnahmen, die das Scraping gerade verhindern sollen, verletzt den Betreiber der Webseite oder Datenbank in seinem Recht, das Publikum seiner Inhalte selbst bestimmen zu dürfen (von Schönfeld 2018: 356 f.).[38] Dabei sind mit „allgemein zugänglichen" Daten nicht nur frei und unmittelbar verfügbare Informationen gemeint, sondern auch solche, die erst nach Zahlung einer Gebühr oder eines Entgelts eingesehen werden können (Schulze-Fielitz 2013: Art. 5 Rn. 80; von Schönfeld 2018: 357). Dieses Kriterium wurde von der höchstrichterlichen Rechtsprechung in ihren Urteilen zum Web Scraping aus urheber- und wettbewerbsrechtlicher Sicht bereits ebenso statuiert, als die Überwindung technischer Schutzmaßnahmen einerseits einen Lauterkeitsverstoß darstellt[39] und dies andererseits durch § 95a UrhG verboten wird.[40] Hintergrund ist die Idee, dass jeder, der sich die öffentliche Zugänglichkeit des Internets für seine Dienste zunutze macht, grundsätzlich auch solche Zugriffe auf seine Inhalte akzeptieren muss, die eben diese öffentliche Zugänglichkeit für sich fruchtbar machen (von Schönfeld 2018: 357).

Aufgrund der durch den EuGH neu eröffneten Diskussion um die urheberrechtliche Zulässigkeit von Scraping-Verfahren ist die Rechtslage bis zu einer Grundsatzentscheidung des BGH oder der Schaffung einer allgemeingültigen Vorschrift zum Web Scraping durch den Gesetzgeber nicht abschließend geklärt. Ob Scraper unzulässigerweise in das Datenbankherstellerrecht nach § 87b UrhG eingreifen, wurde in den oben dargestellten Gerichtsentscheidungen immer anhand der konkreten Umstände des jeweiligen Einzelfalls bewertet. Gewisse Rechtsklarheit und -sicherheit bietet für Zwecke der wissenschaftlichen Forschung aber der neu geschaffene § 60d UrhG, der eine neue Schrankenregelung für Text und Data Mining eingeführt hat und damit selbst bei Bejahung eines Eingriffs in die Rechte des Datenbankherstellers dafür sorgt, dass dieser den Datenzugriff hinnehmen muss. Der „scrapende" Forschende muss allerdings darauf achten, alle Voraussetzungen des § 60d UrhG einzuhalten, insbesondere das Korpus und das Ursprungsmaterial nach Abschluss des Forschungsprojekts zu löschen. Zudem hat der Rechteinhaber (etwa der Datenbankhersteller) gemäß § 60h Abs. 1 UrhG Anspruch auf Zahlung einer angemessenen Vergütung (kritisch dazu Raue 2017: 656 (661 f.)). Der Vergütungsanspruch kann gemäß § 60h Abs. 4 UrhG nur durch eine Verwertungsgesellschaft geltend gemacht werden, der Rechteinhaber selbst ist nicht zur Forderungseinziehung befugt (Hagemeier 2018: § 60h Rn. 11). Nach § 60h Abs. 5 S. 1 UrhG ist ein im Rahmen einer Einrichtung Forschender nicht selbst Schuldner der Vergütung, sondern die Einrichtung als solche. Die Vergütung kann dabei als Pauschale entrichtet werden, möglich ist gemäß § 60h Abs. 3 S. 1 UrhG aber auch eine nutzungsabhängige Vergütung auf Basis repräsentativer Stichproben. Das Wahlrecht dürfte dabei dem Vergütungsschuldner, also dem Forschenden resp. seiner Einrichtung zustehen, wobei sich in der Praxis in aller Regel die Vertragspartner in Vergütungsverhandlungen über Höhe und Art der Vergütung verständigen werden (Pflüger und Hinte 2018: 153 (157). Dadurch, dass die Vergütungsschuld von Gesetzes wegen entsteht, hat der Forschende von sich aus den Rechte-

38 *BVerfG*, Urteil v. 24.01.2001 – 1 BvR 2623/95 u. 622/99 – Fernsehaufnahmen im Gerichtssaal II = BVerfGE 103, 44 = ZUM 2001, 220 (224 f.).

39 Vgl. nur BGH, Urteil v. 30.4.2014 – I ZR 224/12 – Flugvermittlung im Internet = MMR 2014, 740; BGH, Urteil v. 22.06.2011 – I ZR 159/10 – Automobil-Onlinebörse = NJW 2011, 3443.

40 Siehe hierzu oben B. II. 3.

inhaber zu kontaktieren und die Modalitäten der Vergütung zu regeln, wenngleich das Gesetz eine ausdrückliche Informationspflicht nicht normiert.

Aus juristischer Perspektive ist schließlich noch zu beachten, dass durch den Einsatz des Scraping keine technische Schädigung beim Betreiber der Webseite oder Datenbank eintritt (so auch von Schönfeld 2018: 358). Hier lassen sich allerdings keine konkreten Grenzwerte festlegen, ab wann ein Datenabruf schädigend ist. Das hängt letztlich entscheidend von der Rechenkapazität des Hostservers und der technischen Ausgestaltung der Scraping-Software ab. Spätestens bei einer (auch nur vorübergehenden) Funktionsunfähigkeit des Hostservers wegen übermäßiger Anfragen wird allerdings von einer beeinträchtigenden Nutzung auszugehen sein, auf die der Betreiber mit einem Ausschluss der anfragenden IP-Adresse(n) unter Berufung auf sein virtuelles Hausrecht reagieren kann.[41]

Zusammengefasst sind mithin folgende Kriterien zu berücksichtigen:

- Auszuwertende Informationen müssen allgemein zugänglich sein. „Allgemein zugänglich" sind auch solche Daten, die erst nach Zahlung eines Entgelts abgerufen werden können.
- Die Überwindung von technischen Schutzmaßnahmen, die ein Web Scraping verhindern sollen, verletzt den Berechtigten in seinem Recht, das Publikum seiner Inhalte selbst auswählen zu dürfen.
- § 60d UrhG schafft Klarheit hinsichtlich der Zulässigkeit des Web Scrapings für Zwecke der wissenschaftlichen Forschung. Dessen Voraussetzungen sind aber zwingend zu berücksichtigen, etwa:
  - Die wissenschaftliche Forschung darf ausschließlich nicht-kommerziellen Zwecken dienen.
  - Nach Abschluss der Forschungsarbeiten muss das erstellte Korpus gelöscht werden, eine Übermittlung an eine privilegierte Institution (z.B. Bibliothek) ist aber gestattet.
  - Der Rechteinhaber hat Anspruch auf Zahlung einer angemessenen Vergütung.
- Durch den Einsatz von Scraping-Technologien darf keine technische Schädigung beim Betreiber der Webseite oder Datenbank eintreten.

II. Bedingungen für die Archivierung und langfristige Zurverfügungstellung der gewonnenen Daten

1. Urheberrechtliche Anforderungen

Aus urheberrechtlicher Sicht ist wie oben dargestellt eine Löschung des Korpus und der Vervielfältigungsstücke des Ursprungsmaterials nach Abschluss der Forschungsarbeiten vorgeschrieben (§ 60d Abs. 3 S. 1 UrhG). Dabei gibt das Gesetz Art und Umfang der Löschung nicht explizit vor, gemeint ist aber eine unwiederbringliche Beseitigung, etwa durch Zerstörung der Speichermedien oder Löschung der digitalen Datensätze (Hagemeier 2018: § 60d Rn. 19). Empfehlenswert ist die Ausarbeitung eines Löschkonzepts, das etwa eine Beschreibung des Vorgehens zur Einhaltung der gesetzlichen Anforderungen und technische und organisatorische Anforderungen enthält, um den Forschenden vor einer unzulässigen Weiterverwendung der Daten (durch Dritte oder sich selbst) zu schützen und die Befolgung der Löschpflicht zu dokumentieren (ebd.). Wann die Forschungsarbeiten im Sinne des Gesetzes als abgeschlossen gelten, ist schwierig zu bestimmen. Dabei kann der bloße Abschluss des Forschungsberichts noch nicht das Ende markieren, da gemäß § 60d Abs. 1 S. 1 Nr. 2 UrhG ausdrücklich auch Handlungen zur Qualitätskontrolle (etwa Peer-Reviewing) erlaubt sind (Dreier 2018: § 60d Rn. 12).

Erlaubt ist gemäß § 60d Abs. 3 S. 2 UrhG aber eine Übermittlung des Korpus und der Vervielfältigungen des Ursprungsmaterials an Bibliotheken, Archive, Museen und Bildungseinrichtungen (privilegierte Institutionen im Sinne der §§ 60e und 60f UrhG), um die Überprüfung der Einhaltung wissenschaftlicher Standards sowie die Zitier- und Referenzierbarkeit langfristig zu gewährleisten.[42] Dabei setzen Sinn und Zweck der Vorschrift voraus, dass die archivierenden Einrichtungen das übermittelte Material zu Zwecken des Text und Data Mining anderen Forschenden wieder zur Verfügung stellen dürfen; anderenfalls wäre die Aufbewahrung sinnlos (Raue 2017: 656 (661)).

41 Dazu oben B. III. 2.
42 BT-Drs. 18/12329, S. 41.

Eine etwaige elektronische Übermittlung einer archivierenden Bibliothek an andere Forschende ist als öffentliche Wiedergabe im Sinne des § 15 Abs. 2 UrhG zu qualifizieren, die zunächst nicht von den Schrankenregelungen der §§ 60a ff. UrhG erfasst ist (ebd.). Eine entsprechende Berechtigung muss sich aber aus der Vorschrift des § 60d Abs. 3 S. 2 UrhG ergeben, um den Willen des Gesetzgebers angemessen zur Geltung zu bringen (ebd.). Das widerspricht auch nicht der dem § 60d UrhG zugrundeliegenden Urheberrechtsrichtlinie 2001/29/EG[43] (sog. InfoSoc-Richtlinie), die in Art. 5 Abs. 3 lit. a die öffentliche Wiedergabe zu Zwecken der wissenschaftlichen Forschung erlaubt (Raue 2017: 656 (661)). Sicherheitshalber sollten die archivierenden Einrichtungen den Zugang aber auf solche Personen beschränken, die die Voraussetzungen des § 60d Abs. 1 UrhG erfüllen (das empfiehlt Raue 2017: 656 (661)).

2. Datenschutzrechtliche Anforderungen

Datenschutzrechtliche Fragestellungen ergeben sich dann, wenn in dem „gescrapten" Material personenbezogene Daten enthalten sind. Gemäß Art. 4 Nr. 1 DSGVO handelt es sich bei personenbezogenen Daten um „alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden ‚betroffene Person') beziehen". Das ist vor allem dann relevant, wenn Daten aus sozialen Netzwerken ausgewertet werden sollen. Aufgrund des im Datenschutzrecht geltenden Grundsatzes des Verbots mit Erlaubnisvorbehalt ist zunächst jeder Umgang mit personenbezogenen Daten verboten, sofern nicht eine Rechtsnorm die Verarbeitung erlaubt oder der Betroffene in den Datenumgang eingewilligt hat (Art. 6 Abs. 1 DSGVO) (Ingold 2017: Art. 7 Rn. 8 f). Ist Letzteres nicht der Fall, kommt im Kontext des Web Scraping als Erlaubnisnorm in der Regel Art. 6 Abs. 1 lit. f DSGVO in Betracht, wonach die Verarbeitung zulässig ist, wenn nach Abwägung der widerstreitenden Interessen das Forschungsinteresse des Verarbeitenden das Datenschutzinteresse der betroffenen Person überwiegt. Das ist allerdings immer eine Frage des Einzelfalls, für die sich keine allgemeingültigen Kriterien formulieren lassen.

Speziell für den Kontext der wissenschaftlichen Forschung schafft § 27 BDSG n.F. eine Rechtsgrundlage für die Datenverarbeitung zu wissenschaftlichen Forschungszwecken im Hinblick auf besondere Kategorien personenbezogener Daten im Sinne des Art. 9 Abs. 1 DSGVO. Bei Vorliegen der Voraussetzungen (Interessenabwägung und Ergreifen von Maßnahmen zur Wahrung der Interessen der Betroffenen) können insbesondere Gesundheits- und genetische Daten entgegen des absoluten Verbots in Art. 9 Abs. 1 DSGVO auch ohne Einwilligung verarbeitet werden.

Zusätzlich existiert in § 28 BDSG n.F. eine besondere Rechtsgrundlage für Datenverarbeitungen zu im öffentlichen Interesse liegenden Archivzwecken, die gleichzeitig einige Betroffenenrechte einschränkt, um den grundsätzlichen Zweck von Archiven nicht leerlaufen zu lassen (Pauly 2018: § 28 BDSG Rn. 2). Ebenso wie § 27 BDSG n.F. bezieht sich die Vorschrift nur auf besondere Kategorien personenbezogener Daten im Sinne des Art. 9 Abs. 1 DSGVO wie etwa Gesundheitsdaten. Gegenstand von Archivzwecken ist nach Erwägungsgrund 158 der DSGVO insbesondere das Erwerben, Erhalten und Bereitstellen von „Aufzeichnungen von bleibendem Wert für das allgemeine öffentliche Interesse". In allen Fällen muss jedoch überprüft werden, ob der Archivzweck nicht auch mit anonymisierten oder zumindest pseudonymisierten Daten erreicht werden kann (Art. 89 Abs. 1 S. 4 DSGVO) (Pauly 2018: § 28 BDSG Rn. 6).

III. Exkurs: Praktisches Anwendungsbeispiel

Als praktisches Anwendungsbeispiel sei an dieser Stelle auf einzelne Rechtsfragen rund um das Web Scraping von Social-Media-Webseiten (insbesondere Twitter) einzugehen.

1. Bindungswirkung der Nutzungsbedingungen der Twitter-API

Möchten Forschende zum automatisierten Auslesen und Auswerten von Tweets die Twitter-API verwenden, muss vorher deren Nutzungsbedingungen[44] zugestimmt werden. Wie oben (B. III. 1.) dargestellt, kommt ein Nutzungsvertrag noch nicht durch den bloßen Aufruf einer Webseite zustande.

43 Richtlinie 2001/29/EG des Europäischen Parlaments und des Rates vom 22. Mai 2001 zur Harmonisierung bestimmter Aspekte des Urheberrechts und der verwandten Schutzrechte in der Informationsgesellschaft, ABl. 2001 Nr. L 167, 10 ff.

44 https://developer.twitter.com/en/developer-terms/agreement-and-policy (24.08.2018).

Durch das Erfordernis einer ausdrücklichen Zustimmung kommt zwischen dem Verwender und Twitter ein Nutzungsvertrag zustande, der ersteren vertraglich zur Einhaltung der Developer Agreement and Policy verpflichtet. Gegenstand der Developer Policy ist unter anderem die Verpflichtung, die Kontrolle und Privatsphäre des Nutzenden über seine Inhalte zu respektieren. Dazu gehört auch die Pflicht, Kopien von gelöschten Tweets innerhalb eines kurzfristigen Zeitraums zu vernichten.[45] Das ist ungünstig für Forschende: Förderbedingungen verpflichten sie häufig dazu, ihre Forschungsdaten zum Zwecke der Replikation für mehrere Jahre zu archivieren. Bei diesen Policies handelt es sich um Allgemeine Geschäftsbedingungen (AGB), die sich hinsichtlich ihrer Wirksamkeit an den §§ 305 ff. BGB messen lassen müssen (s.o. B. III. 1.). Da Nutzende den Bedingungen ausdrücklich zustimmen muss, wurden sie wirksam in den Nutzungsvertrag einbezogen (§ 305 Abs. 2 BGB). Eine unangemessene Benachteiligung von Nutzenden dadurch, dass sie Kopien von gelöschten Tweets vernichten müssen, ist nach summarischer Prüfung nicht erkennbar. Von einer Wirksamkeit der Bestimmungen ist daher auszugehen.

Allerdings sind die Bedingungen des Nutzungsvertrags mit Twitter bei Verwendung der API für den Vertragspartner bindend. Möchten Forschende ihnen nicht unterworfen sein, bleibt ihnen nur die Möglichkeit, – sicherheitshalber – einen gegebenenfalls vorher abgeschlossenen Nutzungsvertrag zu kündigen und die Daten ohne Verwendung der Programmierschnittstelle zu scrapen. Auch Zweckbegrenzungen, die in den Twitter-API festgelegt sind, binden nur die Vertragsparteien. Nutzende, die sich diesen Bedingungen nicht durch eine auf einen Vertragsschluss gerichtete Abgabe einer entsprechenden Willenserklärung unterworfen haben, sind nicht an sie gebunden. Allerdings können deliktische Unterlassungs- oder Schadensersatzansprüche entstehen, wenn durch das Scraping die Funktionsfähigkeit des Servers beeinträchtigt wird oder es zu anderweitigen Störungen des Betriebsablaufs kommt.

2. Anwendbares Recht bei Auslandsbezug

Möchten in Deutschland arbeitende Forschende auf einer US-amerikanischen Webseite Scraping-Verfahren anwenden, um dadurch soziale Gruppen in einem Drittland zu untersuchen, stellt sich die Frage nach dem für sie maßgeblichen Rechtssystem. Grundsätzlich sind Forschende in diesem Fall zunächst einmal dem deutschen Rechtssystem unterworfen, da sie sich innerhalb von dessen räumlichem Anwendungsbereich aufhalten. Schwierigkeiten bereitet das Urheberrecht: Nach dem sog. Schutzlandprinzip ist für Fragen des geistigen Eigentums das Recht desjenigen Staates anwendbar, für dessen Gebiet der Schutz begehrt wird.[46] Bei der Frage nach dem anwendbaren Recht bei Handlungen im Internet ergeben sich allerdings erhebliche Unwägbarkeiten, da gerade bei grenzüberschreitenden Datentransfers mehrere Rechtsordnungen berührt werden und die Zuordnung eines Verhaltens zum Territorium eines konkreten Staates daher kompliziert ist (anschaulich Bollacher 2005: 101 ff.). So könnte nach dem Schutzlandprinzip im o.g. Sachverhalt auch US-amerikanisches Urheberrecht zu beachten sein, wenn die Server, auf denen die auszulesenden Datenbanken liegen, im Territorium der USA stehen. Folge des Schutzlandprinzips ist demnach, dass der Inhaber eines nach US-amerikanischem Recht bestehenden Urheberrechts Ansprüche nach dieser Rechtsordnung gegen den Scraper geltend machen könnte, wenn dieser bei seiner Tätigkeit Urheber- oder Schutzrechte verletzt. Voraussetzung dafür ist aber, dass der Handlungsort der Urheberrechtsverletzung in den USA zu sehen ist. Ob im konkreten Anwendungsfall des Web Scraping der Handlungsort am Standort des Servers (hier: USA) oder am Standort des Scrapenden (hier: Deutschland) zu sehen ist, wurde – soweit ersichtlich – noch nicht untersucht.

Andere Rechtssysteme können außerdem dann beachtlich werden, wenn der Forschende Partei eines Vertrags wird, der eine Klausel bezüglich des anwendbaren Rechts enthält. Diese Bestimmung bezieht sich dann allerdings nur auf solche Streitigkeiten oder Sachverhalte, die sich unmittelbar aus dem jeweiligen Vertrag ergeben. Im Übrigen bleibt es bei der maßgeblichen Anwendbarkeit deutschen Rechts (dazu ausführlich oben B. II. 3. und B. III.).

---

45 Twitter Developer Policy, I. Guiding Principles, C. Respect Users' Control and Privacy 3., abrufbar unter https://developer.twitter.com/en/developer-terms/agreement-and-policy (24.08.2018).

46 So geregelt in Art. 5 Abs. 2 S. 2 der Berner Übereinkunft zum Schutz von Werken der Literatur und Kunst, aufgegriffen in Art. 9 des TRIPS-Abkommens (Übereinkommen über handelsbezogene Aspekte der Rechte des geistigen Eigentums) und in Art. 8 Abs. 1 der Rom-II-VO; vgl. *Rehbinder/Peukert*, Urheberrecht, § 49 Rn. 1206.

3. Datenschutzrechtliche Implikationen

a) Öffentliches Interesse im Sinne des Art. 89 DSGVO

Art. 89 Abs. 1 DSGVO privilegiert neben wissenschaftlichen und historischen Forschungszwecken auch „im öffentlichen Interesse liegende Archivzwecke". Unter Archiven versteht die DSGVO ausweislich ihres Erwägungsgrundes 158 S. 2 „Behörden oder öffentliche oder private Stellen, die Aufzeichnungen von öffentlichem Interesse führen". Erfasst sind aber nur solche Archivzwecke, an denen ein öffentliches Interesse besteht. Als Beispiele für solche nennt Erwägungsgrund 158 S. 4 „die Bereitstellung spezifischer Informationen im Zusammenhang mit dem politischen Verhalten unter ehemaligen totalitären Regimen, Völkermord, Verbrechen gegen die Menschlichkeit, insbesondere dem Holocaust, und Kriegsverbrechen". Entscheidend ist, dass die Archivierung nicht nur den Interessen des Datenverarbeiters, sondern zugleich auch der Gesellschaft insgesamt dient (Schantz 2017: Rn. 1347).

Gleichzeitig erlauben Art. 89 Abs. 2 und 3 DSGVO, gesetzliche Ausnahmen von den Betroffenenrechten der Art. 12 ff. DSGVO einzuführen. So muss unter anderem sogar das Recht auf Löschung („Recht auf Vergessenwerden") nach Art. 17 DSGVO zurücktreten, wenn sich eine längere Speicherung der Daten zur Sicherung des Forschungszwecks als erforderlich herausstellt (Hense 2017: Art. 89 Rn. 12). Das ergibt sich nicht aus Art. 89 DSGVO, sondern bereits unmittelbar aus den Ausnahmen zum Recht auf Löschung aus Art. 17 Abs. 3 lit. d sowie Art. 5 Abs. 1 lit. e DSGVO (vgl. auch Schantz 2017: Rn. 1361).

b) Personenbeziehbarkeit bei Pseudonymisierung

Abschließend stellt sich noch die Frage, wann trotz ausschließlicher Alias-Nennung (Pseudonymisierung) Daten als personenbeziehbar gelten und damit gemäß Art. 2 Abs. 1 DSGVO dem sachlichen Anwendungsbereich des Datenschutzrechts unterfallen. Art. 4 Nr. 5 DSGVO definiert Pseudonymisierung als „Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden". Im Gegensatz zu vollständig anonymisierten Daten, bei denen ein Personenbezug nicht (mehr) herstellbar ist und die damit nicht dem Anwendungsbereich der DSGVO unterfallen, handelt es sich bei pseudonymisierten Daten weiterhin um potentiell personenbezogene Daten, weshalb sie dem Grunde nach vollständig der DSGVO unterliegen (vgl. Erwägungsgrund 26 S. 2) (Schantz 2017: Rn. 303).

Stimmen in der rechtswissenschaftlichen Fachliteratur schränken diese Absolutheit aber dann ein, wenn die Wahrscheinlichkeit einer Bestimmung der hinter einem Alias stehenden Person so gering ist, dass das Risiko praktisch vernachlässigbar ist (Roßnagel 2018: 243 (244)). Eine Person ist nämlich nur dann als identifizierbar anzusehen, wenn die Bestimmbarkeit faktisch durchführbar ist.[47] Die Wahrscheinlichkeit einer Zuordnung ergibt sich aus einer Mittel-Zweck-Abwägung, wobei hierbei nach Erwägungsgrund 26 S. 4 die „Kosten der Identifizierung und der dafür erforderliche Zeitaufwand" und „die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen" sind (vgl. auch Roßnagel 2018: 243 (244)). Das ist letztlich eine Frage des jeweiligen Einzelfalls, für die sich keine verallgemeinerungsfähigen Kriterien aufstellen lassen. Hat ein Forschender überhaupt keine Möglichkeit, das Pseudonym einer konkreten natürlichen Person zuzuordnen, handelt es sich für ihn nicht um personenbezogene Daten mit der Folge, dass er sich bei der Verarbeitung dieser Daten nicht an die Vorgaben der DSGVO und des BDSG halten muss.

47 *EuGH*, Urteil v. 19.10.2016 – C-582/14, Rn. 46 – *Breyer* = NJW 2016, 3579.

D. Fazit und Ausblick

Abschließend lässt sich festhalten, dass die juristische Bewertung des Web Scraping vor allem im Urheberrecht erheblichen Erörterungsbedarf bereitet. Nachdem die höchstrichterliche Rechtsprechung des BGH und des EuGH auf keinen gemeinsamen Nenner gekommen ist und auch die – zeitlich nachfolgende – europäische Rechtsprechung ausdrücklich keine allgemeingültigen Aussagen betreffend die Zulässigkeit des Web Scraping treffen wollte, war die Rechtslage von erheblicher Rechtsunsicherheit geprägt. Mit dem UrhWissG und der neu geschaffenen Vorschrift des § 60d UrhG hat der Gesetzgeber aber zumindest für den Bereich der – im vorliegenden Gutachten maßgeblichen – wissenschaftlichen Forschung diesen Missstand beseitigt und gewisse Rechtssicherheit geschaffen.

Damit ist er bewusst der Europäischen Union vorweggeeilt, die derzeit über den Entwurf einer Richtlinie über das Urheberrecht im digitalen Binnenmarkt berät.[48] An Art. 3 des Richtlinienentwurfs orientiert sich die Neuregelung des § 60d UrhG. Aufgrund der Richtlinie sah sich der deutsche Gesetzgeber verpflichtet, die Nutzung von Text und Data Mining auf die wissenschaftliche Forschung zu nicht-kommerziellen Zwecken zu begrenzen; eine Ausweitung der Erlaubnis auf forschende Unternehmen ist aufgrund der europäischen Vorgaben nicht absehbar (Hagemeier 2018: § 60d Rn. 4). In der Gesetzesbegründung zum UrhWissG hat der Gesetzgeber bereits angekündigt, die Vorschrift nötigenfalls an die Vorgaben der Richtlinie anzupassen, sobald sie in Kraft tritt.[49]

48 Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates über das Urheberrecht im digitalen Binnenmarkt, COM(2016) 593 final, abrufbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52016PC0593 (24.08.2018).

49 BT-Drs. 18/12329, S. 40.

Literaturverzeichnis

Bitkom (2013): Trends im E-Commerce. Konsumverhalten beim Online-Shopping. Berlin, Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. (BITKOM). https://www.bitkom.org/Publikationen/2013/Studien/Trends-im-E-Commerce/BITKOM_E-Commerce_Studienbericht.pdf (Zugriff am 16.08.2018).

Bitkom (2018): Markt für Big Data wächst in Deutschland zweistellig. Presseinformation vom 13.03.2018. Berlin, Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. (BITKOM). https://www.bitkom.org/Presse/Presseinformation/Markt-fuer-Big-Data-waechst-in-Deutschland-zweistellig.html (Zugriff am 16.08.2018).

Bollacher, Philipp D. (2005): Internationales Privatrecht, Urheberrecht und Internet. Frankfurt am Main, Verlag Peter Lang.

Deutsch, Askan (2009): Die Zulässigkeit des so genannten „Screen-Scraping" im Bereich der Online-Flugvermittler. GRUR (Gewerblicher Rechtsschutz und Urheberrecht) 2009(11), 1027–1032.

Dreier, Thomas (2018): Kommentierung des § 60d UrhG sowie der Vorbemerkung zu §§ 87a ff. UrhG. In: Thomas Dreier und Gernot Schulze (Hrsg): Urheberrechtsgesetz. Kommentar. 6. Auflage, München, C.H.BECK.

Hagemeier, Stefanie (2018): Kommentierung der §§ 60d, 60f, 60h UrhG. In: Hartwig Ahlberg und Horst-Peter Götting (Hrsg.): Beck'scher Online-Kommentar zum Urheberrecht. 21. Edition (Stand: 04.06.2018), München, C.H.BECK.

Hense, Ansgar (2017): Kommentierung des Art. 89 DSGVO. In: Gernot Sydow (Hrsg.): Europäische Datenschutzgrundverordnung. Handkommentar. Baden-Baden, Nomos.

Ingold, Albert (2017): Kommentierung des Art. 7 DSGVO. In: Gernot Sydow (Hrsg.): Europäische Datenschutzgrundverordnung. Handkommentar. Baden-Baden, Nomos.

Kinne, Jan und Janna Axenbeck (2018), Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany, ZEW Discussion Paper No. 18-033, Mannheim.

Kotthoff, Jost (2013): Kommentierung der §§ 4 und 87a UrhG. In: Gunda Dreyer, Jost Kotthoff und Astrid Meckel: Urheberrecht. Kommentar. 3. Auflage, Heidelberg u.a., C.F. Müller.

Kukulenz, Dirk (2008): Die Dynamik des World Wide Web und Konsequenzen für die Informationssuche. Habilitationsschrift. https://www.ifis.uni-luebeck.de/fileadmin/user_files/veroeffentlichungen/Habil-Kukulenz08.pdf (Zugriff am 16.08.2018).

Leupold, Andreas und Dominik Demisch (2000): Bereithalten von Musikwerken zum Abruf in digitalen Netzen. ZUM (Zeitschrift für Urheber- und Medienrecht) 2000(5), 379–390.

Loewenheim, Ulrich (2017): Kommentierung des § 4 UrhG. In: Gerhard Schricker und Ulrich Loewenheim (Hrsg.): Urheberrecht. Kommentar. 5. Auflage, München, C.H.BECK.

Marquardt, Malte (2014): Kommentierung des § 4 UrhG. In: Artur-Axel Wandtke und Winfried Bullinger (Hrsg.): Praxiskommentar zum Urheberrecht. 4. Auflage, München, C.H.BECK.

Maume, Philipp (2007): Bestehen und Grenzen des virtuellen Hausrechts. MMR (MultiMedia und Recht) 2007(10), 620–625.

Mörike, Matthias (2018): Im Namen der Wissenschaft! Zur Zulässigkeit von Screen Scraping im Forschungsbetrieb vor dem Hintergrund des neuen Urheberrechts. DFN-Infobrief 6/2018. https://www.dfn.de/fileadmin/3Beratung/Recht/1infobriefearchiv/2018/Infobrief_Recht_06-2018.pdf (Zugriff am 16.08.2018).

Nordemann, Axel (2014): Einleitung zum Urheberrechtsgesetz. In: Friedrich Karl Fromm und Wilhelm Nordemann (Begr.): Urheberrecht. Kommentar zum Urheberrechtsgesetz, Verlagsgesetz, Urheberrechtswahrnehmungsgesetz. 11. Auflage, Stuttgart, Kohlhammer.

Pauly, Daniel A. (2018): Kommentierung des § 28 BDSG. In: Boris P. Paal und Daniel A. Pauly (Hrsg.): Datenschutz-Grundverordnung und Bundesdatenschutzgesetz. Kommentar. 2. Auflage, München, C.H.BECK.

Pflüger, Thomas und Oliver Hinte (2018): Das Urheberrechts-Wissensgesellschafts-Gesetz aus Sicht von Hochschulen und Bibliotheken. ZUM (Zeitschrift für Urheber- und Medienrecht) 2018(3), 153–161.

Raue, Benjamin (2017): Text und Data Mining. Die neue Urheberrechtsschranke des § 60d UrhG. CR (Computer und Recht) 2017(10), 656–662.

Redeker, Helmut (2007): Anmerkung zu LG München I, Urteil v. 25.10.2006 – 30 O 11973/05. CR (Computer und Recht) 2007(4), 265–267.

Rehbinder, Manfred und Alexander Peukert (2018): Urheberrecht und verwandte Schutzrechte. 18. Auflage, München, C.H.BECK.

Roßnagel, Alexander (2018): Pseudonymisierung personenbezogener Daten. Ein zentrales Instrument im Datenschutz nach der DS-GVO. ZD (Zeitschrift für Datenschutz) 2018(6), 243–247.

Schack, Haimo (2001): Urheberrechtliche Gestaltung von Webseiten unter Einsatz von Links und Frames. MMR (MultiMedia und Recht) 2001(1), 9–17.

Schantz, Peter und Heinrich Amadeus Wolff (2017): Das neue Datenschutzrecht. Datenschutz-Grundverordnung und Bundesdatenschutzgesetz in der Praxis. München, C.H.BECK.

Schapiro, Leo und Konrad Żdanowiecki (2015): Screen Scraping. Rechtlicher Status quo in Zeiten von Big Data. MMR (MultiMedia und Recht) 2015(8), 497–501.

Schulze-Fielitz, Helmuth (2013): Kommentierung des Art. 5 GG. In: Horst Dreier (Hrsg.): Grundgesetz Kommentar, Band I: Präambel, Artikel 1-19. 3. Auflage, Tübingen, Mohr Siebeck.

Seagate (n.d.): Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2016 und 2025 (in Zettabyte). Statista - Das Statistik-Portal. https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/ (Zugriff am 16.08.2018).

Thum, Dorothee und Kai Hermes (2014): Kommentierung des § 87a UrhG. In: Artur-Axel Wandtke und Winfried Bullinger (Hrsg.): Praxiskommentar zum Urheberrecht. 4. Auflage, München C.H.BECK.

Vogel, Martin (2017): Kommentierung des § 87a UrhG. In: Gerhard Schricker und Ulrich Loewenheim (Hrsg.): Urheberrecht. Kommentar. 5. Auflage, München, C.H.BECK.

Von Schönfeld, Max (2018): Screen Scraping und Informationsfreiheit. Baden-Baden, Nomos.

Impressum

Herausgeberschaft:
Rat für Sozial- und Wirtschaftsdaten (RatSWD)
Rungestr. 9
10179 Berlin
office@ratswd.de
www.ratswd.de

Redaktion:
Thomas Runge, Dr. Nora Dörrenbächer, Dr. Mathias Bug, Lea Salathé

Gestaltung/Satz:
Claudia Kreutz

Icons:
made by Freepik from www.flaticon.com
Font Awesome, fontawesome.com (angepasst)

Berlin, November 2019

RatSWD Output:
Die RatSWD Output Series dokumentiert die Arbeit des RatSWD in seiner 6. Berufungsperiode (2017–2020). In ihr werden seine Stellungnahmen und Empfehlungen veröffentlicht und auf diesem Weg einer breiten Leserschaft zugänglich gemacht.

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01UW1802 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt, sofern nicht anders ausgewiesen, beim RatSWD.

doi: 10.17620/02671.39

Zitationsvorschlag:

RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2019): Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement. RatSWD Output 4 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.39.

Der Rat für Sozial- und Wirtschaftsdaten (RatSWD) berät seit 2004 die Bundesregierung und die Regierungen der Länder in Fragen der Forschungsdateninfrastruktur für die empirischen Sozial-, Verhaltens- und Wirtschaftswissenschaften. Im RatSWD arbeiten acht durch Wahl legitimierte Vertreterinnen und Vertreter der sozial-, verhaltens- und wirtschaftswissenschaftlichen Fachdisziplinen mit acht Vertreterinnen und Vertretern der wichtigsten Datenproduzenten zusammen.

Er versteht sich als institutionalisiertes Forum des Dialoges zwischen Wissenschaft und Datenproduzenten und erarbeitet Empfehlungen und Stellungnahmen. Der RatSWD engagiert sich für eine Infrastruktur, die der Wissenschaft einen breiten, flexiblen und sicheren Datenzugang ermöglicht. Solche Daten werden von staatlichen, wissenschaftsgetragenen und privatwirtschaftlichen Akteuren bereitgestellt. Der RatSWD hat 34 Forschungsdatenzentren akkreditiert, deren Kooperationen er fördert.

www.ratswd.de