Handreichung zur rechtskonformen Durchführung von Web-Scraping Projekten in der nicht-kommerziellen wissenschaftlichen Forschung[1]

Das (Web-) Scraping (häufig auch mit Methoden des (Web-) Crawling verbunden) ist ein Verfahren um automatisiert bereitgestellte (Web-) Inhalte zu verarbeiten und in eine maschinenlesbare Form zu überführen. Um Klarheit hinsichtlich der Frage inwieweit die Forschung diese Methoden verwenden darf zu gewinnen, hat der Rat für Sozial- und Wirtschaftswissenschaften(RatSWD) ein Rechtsgutachten bei der Forschungsstelle RobotRecht der Universität Würzburg in Auftrag gegeben. Hierin werden zentrale Kriterien für den Einsatz der Methode vorgestellt (RatSWD, 2019, S. 45ff.):

- „Auszuwertende Informationen müssen allgemein zugänglich sein. [Dies beinhaltet] auch solche Daten, die erst nach Zahlung eines Entgelts abgerufen werden können." (ebd., S. 45f.)
- Überwindung von technischen Schutzmaßnahmen (bspw. Verbot in robots.txt oder Captcha) ist eine Verletzung von Betreiberrechten und damit unzulässig (vgl. ebd., S. 45)
- Zwingende Beachtung der Voraussetzungen für UrhG §60d:
    - „Die wissenschaftliche Forschung darf ausschließlich nicht-kommerziellen Zwecken dienen." (vgl. ebd., S. 46).
    - Löschung der Daten nach Abschluss der Forschungsarbeiten; Übermittlung an privilegierte Institutionen (wissenschaftliche Archive, Bibliotheken) gestattet (vgl. ebd.).
    - „Der Rechteinhaber hat Anspruch auf Zahlung einer angemessenen Vergütung" (Geltendmachung via Verwertungsgesellschaft an die wissenschaftliche Einrichtung) (vgl. ebd., S. 45f).
- „Durch den Einsatz von (Web-)Scraping-Technologien darf keine technische Schädigung beim Betreiber der Website [...] eintreten" (ebd., S. 46).

Darüber hinaus können je nach konkretem Projekt noch datenschutzrechtliche oder vertragsrechtliche Verpflichtungen hinzukommen.

Ratschläge zur praktischen Umsetzung der rechtlichen Vorgaben werden im nachfolgenden für die einzelnen Rechtsbereiche vorgestellt und abschließend durch allgemeine Vorschläge zum Projektmanagement ergänzt.

Virtuelles Hausrecht

Um keine technische Schädigung (beispielsweise temporäre Nichtverfügbarkeit des Dienstes) der bereitgestellten Webseite oder Datenbank hervorzurufen, ist die Limitierung der Anzahl der gestellten Anfragen dringend erforderlich. Konkrete Grenzwerte sind nicht vorgegeben, dies ist vom Einzelfall abhängig zu machen (vgl. ebd., S. 46). Auch sollte darauf geachtet werden, dass keine technischen Lücken, die den Dienst beeinträchtigen könnten, beim (Web-)Scraping verwendet werden (beispielsweise der Abruf großer Teile der Datenbank mit nur einer Anfrage).

[1] Stand: 30.01.2020

Urheberrecht

Die Rechtsgrundlage des (Web-) Scraping bildet die Schrankenreglung im UrhG §60d: Text und Data Mining, welche die Vervielfältigung eigentlich urheberrechtlich-geschützten Materials im Kontext von Text und Data Mining in der nicht-kommerziellen Forschung erlaubt. Wichtig ist neben den oben genannten Voraussetzungen, dass dennoch eine Vervielfältigung außerhalb der Forschungsgruppe und über den Forschungszeitraum hinaus (ausgenommen die privilegierten Institutionen) unter allen Umständen zu vermeiden ist. Bezüglich der Zahlung der angemessenen Vergütung fehlen aktuell die praktischen Abläufe und Verfahrensweisen. Nichtsdestotrotz ist eine Kontaktaufnahme zum/zur Webseiten- Betreiber/in empfehlenswert und möglicherweise sogar erforderlich (vgl. ebd, S45f.).

Darüber hinaus ist es nicht erlaubt, technischen Schutzmaßnahmen zu umgehen. Die für Suchmaschinen auf vielen Webseiten vorhandene Robots.txt-Datei (Pfad: „domain.tld/robots.txt“) ist demnach im Vorhinein bezüglich der abzurufenden Seiten auf entsprechende „Disallow“-Einträge zu überprüfen. Sobald Probleme beim Abrufen der Webseiten (beispielsweise in Form von IP-Sperren oder der Anzeige von Captcha) auftreten, wäre ein Scraping ebenso einzustellen.

Datenschutzrecht

Gegebenenfalls werden Daten, welche „Informationen [enthalten], die sich auf eine identifizierte oder identifizierbare natürliche Person [...] beziehen“ (DSGVO Art.4 Abs.1) verarbeitet. Je nach Anwendungsfall können auch besondere Arten personenbezogener Daten (DSGVO Art. 9) verarbeitet werden (beispielsweise politische Überzeugungen). Daran anschließend wäre die Frage zu stellen, auf welcher Rechtsgrundlage diese Verarbeitung stattfindet. Die Einholung von Einwilligungen in (Web-) Scraping Projekten ist häufig nicht praktikabel. Allerdings können je nach Anwendungsfall Forschungszwecke geltend gemacht werden (Golla, Hofmann, & Bäcker, 2018). Nichtsdestotrotz müsste gegebenenfalls eine Abwägung zwischen diesen Forschungsinteressen und den Persönlichkeitsinteressen der Betroffenen stattfinden (vgl. ebd., S. 91ff. , S.100) sowie (RatSWD, 2019, S. 47).

Insbesondere wenn ein Scraping nicht von einem Computersystem der/des Forschenden ausgeht, sondern mit einem Browser-Plugin oder einer anderen Software, welche auf den Geräten von Versuchspersonen installiert wird, durchgeführt wird, dann werden in jedem Fall personenbezogene Daten verarbeitet. In diesem Fall ist jedoch eine Zustimmung der Versuchspersonen dringend erforderlich und bildet die Rechtsgrundlage zur Verarbeitung der Daten.

Um eine datenschutzkonforme Verarbeitung zu gewährleisten sind folgende Maßnahmen in frühzeitiger Rücksprache mit dem Datenschutzbeauftragten zu treffen (Golla, Hofmann, & Bäcker, 2018, S. 100):

- Klare Darlegung der Forschungsfrage; Erstellung eines Datenmanagmentplans (vgl. ebd.)
- Eintrag in das Verzeichnis von Verarbeitungstätigkeiten (DSGVO Art. 30) sowie die Dokumentation technisch-organisatorischer Maßnahmen (DSGVO Art. 32)[2]

[2] Siehe dazu die im Sharepoint der UHH bereitgestellte Möglichkeit Verarbeitungstätigkeiten anzulegen und die Seiten des Datenschutzbeauftragten der UHH https://www.isdsm.uni-hamburg.de/dokumente.html

- Gewährleistung der Betroffenenrechte sowie informierte freiwillige Einwilligung[3]
- technische Vorkehrungen zur Datenminimierung
- Nutzung von Anonymisierungs- bzw. Pseudoanonymisierungsmöglichkeiten
- Festlegung von Speicherfristen sowie deren Befolgung (ebd.)

Vertragsrecht

Gegebenenfalls sind rechtliche Vereinbarungen getroffen worden, um an bestimmte Inhalte zu gelangen. Beispielsweise zur Benutzung einer Programmierschnittstellen (Twitter-API), oder auch bezüglich des Zugangs zu wissenschaftlichen Publikationen bei Verlagen (häufig über die Bibliotheken). Hier können die Bedingungen des Nutzungsvertrages bestimmte wissenschaftliche Qualitätsansprüche (beispielsweise Speicherung der Daten zur Möglichkeit der Replikation der Forschung) verhindern (RatSWD, 2019, S. 47f.). Es sollte demnach immer eine genaue Analyse der getroffenen Vereinbarungen stattfinden. Falls diese eine Forschung mit dem geforderten Qualitätsanspruch verunmöglichen, sollte geprüft werden, ob ein Datenzugang ohne den Abschluss eines Vertrages möglich ist.

Allgemeine Ratschläge zum Projektmanagement

Generell und auch im Hinblick auf die urheber- und datenschutzrechtlichen Verpflichtungen empfiehlt sich die Führung eines Forschungsdatenmanagementplans. Darüber hinaus ist eine Versionierung der entwickelten Software hilfreich, um auch im Nachhinein jeweils getätigte Aktionen nachvollziehen zu können. Die Universität Hamburg stellt hierfür entsprechende Tools bereit[4].

Literaturverzeichnis

Golla, S. J., Hofmann, H., & Bäcker, M. (2018). Connecting the Dots. Sozialwissenschaftliche Forschung in Sozialen Online-Medien im Lichte von DS-GVO und BDSG-neu. *Datenschutz und Datensicherheit - DuD*(42), S. 89-100. doi:10.1007/s11623-018-0900-x

RatSWD. (2019). *Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement.* Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD). doi:10.17620/02671.39

Watteler, O., & Ebel, T. (2019). Datenschutz im Forschungsdatenmanagement. In U. Jensen, S. Netscher, & K. Weller, *Forschungsdatenmanagement sozialwissenschaftlicher Umfragen. Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (S. 57-80). Obladen, Berlin, Toronto: Verlag Barbara Budrich. doi:10.3224/84742233.05

[3] Siehe hierzu auch (Watteler & Ebel, 2019, S. 57ff.).

[4] Git-Versionsverwaltungssystem: https://gitlab.rrz.uni-hamburg.de; Erstellung Forschungsdatenmanagementplan: https://dmp.fdm.uni-hamburg.de/ sowie weitere Informationen: https://www.fdm.uni-hamburg.de/fdm/datenmanagementplan.html