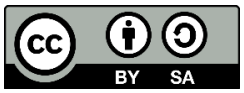
Datentracking in der Wissenschaft: Aggregation und Verwendung bzw. Verkauf von Nutzungsdaten durch Wissenschaftsverlage

Ein Informationspapier des Ausschusses
für Wissenschaftliche Bibliotheken und Informationssysteme
der Deutschen Forschungsgemeinschaft
28. Oktober 2021

Deutsche Forschungsgemeinschaft
Kennedyallee 40 · 53175 Bonn · Postanschrift: 53170 Bonn
Telefon: + 49 228 885-1 · Telefax: + 49 228 885-2777 · postmaster@dfg.de · www.dfg.de

Deutsche Forschungsgemeinschaft e.V.
Kennedyallee 40 · 53175 Bonn
Postanschrift: 53170 Bonn
Telefon: +49 228 885-1
Telefax: +49 228 885-2777
postmaster@dfg.de
www.dfg.de



Oktober 2021

Dr. Angela Holzer
Gruppe Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)
Tel. +49 (228) 885-2568
angela.holzer@dfg.de

Stand: 28. Oktober 2021

DOI: 10.5281/zenodo.5900759

Zitiervorschlag: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme (2021): Datentracking in der Wissenschaft: Aggregation und Verwendung bzw. Verkauf von Nutzungsdaten durch Wissenschaftsverlage. Ein Informationspapier des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme der Deutschen Forschungsgemeinschaft.

Datentracking in der Wissenschaft

Dieses Informationspapier des Ausschusses für wissenschaftliche Literaturversorgungs- und Informationssysteme (AWBI) der Deutschen Forschungsgemeinschaft (DFG) zum Thema Datentracking bei digitalen wissenschaftlichen Ressourcen beschreibt Möglichkeiten der digitalen Nachverfolgung von wissenschaftlichen Aktivitäten. Es legt die Transformation von Wissenschaftsverlagen hin zu Data Analytics Businesses dar, weist auf die Konsequenzen daraus für die Wissenschaft und deren Einrichtungen hin und benennt die zum Einsatz kommenden Typen der Datengewinnung. Damit dient es vor allem der Darstellung gegenwärtiger Praktiken und soll zu Diskussionen und Positionen über deren Konsequenzen für die Wissenschaft anregen. Es richtet sich an alle Akteure in der Wissenschaftslandschaft.

1. Beschreibung der aktuellen Situation

In den letzten Jahren sind digitale Datenmärkte unterschiedlichster Art entstanden, die sich in öffentliche, wissenschaftliche und kommerzielle unterscheiden lassen.[1] Im Bereich der Wissenschaft gibt es neben sehr positiven Entwicklungen, die sich nicht zuletzt in der besseren Aufbereitung, rechtlichen Regelungen zu und Nutzung von Forschungsdaten niederschlagen, auch Entwicklungen, deren Konsequenzen ausführlich reflektiert und gegebenenfalls reguliert werden müssen. Diese Entwicklungen werden im Folgenden dargestellt. Für die Wissenschaft möglicherweise nachteilige Effekte entstehen sowohl durch eine Vermischung von wissenschaftlichen und kommerziellen Bereichen als auch durch Regelungslücken oder unterschiedliche internationale Gesetzeslagen.

Seit einiger Zeit verändern die großen wissenschaftlichen Verlage ihr Geschäftsmodell grundlegend mit erheblichen Auswirkungen auf die Wissenschaften: Die Aggregation und die Weiterverwendung bzw. der Weiterverkauf von Nutzerspuren werden relevante Aspekte der Verlagstätigkeit.[2] Verlage verstehen sich jetzt teilweise ausdrücklich als Unternehmen für Informationsanalysen.[3] Das Geschäftsmodell der Verlage wandelt sich damit von Content Providern hin zu einem Data Analytics Business. Dabei werden die Daten von Wissenschaftlerinnen und Wissenschaftlern (das heißt personalisierte Profile, Zugriffs- und Nutzungsdaten, Verweildauern bei Informationsquellen usw.) bei der Nutzung von Informationsdiensten wie z. B. der Literaturrecherche getrackt, das heißt festgehalten und gespeichert. Wissenschaftstracking erfolgt durch ein Ensemble an Werkzeugen, die vom Nachverfolgen von Seitenbesuchen über

[1] Putnings, M., „Datenmarkt“, in: *Praxishandbuch Forschungsdatenmanagement*, 2021, S. 143, Praxishandbuch Forschungsdatenmanagement (degruyter.com).
[2] Aspesi, C., Allen, N. S., Crow, R., Daugherty, S., Joseph, H., McArthur, J. T., & Shockey, N., *SPARC Landscape Analysis*, 2019, March 29, https://doi.org/10.31229/osf.io/58yhb.
[3] Z. B. Selbstdarstellung von Elsevier: „Elsevier ist ein globales Unternehmen für Informationsanalysen, das Institutionen und Fachleute dabei unterstützt, die Leistungen im Gesundheitswesen und in der Wissenschaft zum Wohle der Menschheit zu verbessern.“ www.elsevier.com/de-de/about.

Authentifizierungssysteme bis zu detaillierten Echtzeitdaten über das Informationsverhalten von Einzelnen und Institutionen reichen. Die Erfassung von u. a. Seitenbesuchen, Zugriffen, Downloads und damit granularer Profile des wissenschaftlichen Verhaltens erfolgt teilweise unter unzureichender Information der Nutzenden. Daten aus verschiedenen Quellen können aggregiert und mit weiteren Informationen über die Personen, auch aus dem nicht wissenschaftlichen Umfeld, kombiniert werden.

Die Funktion dieser Datensammlung durch Verlage ist eine doppelte: Zum einen geht es um die Erschließung eines neuen Geschäftsfeldes, dem Handeln mit Daten über Wissen, wissenschaftliche Entwicklungen und ihre Akteure. Zum anderen geht es um den Ausbau der Dienstleistungen der großen Wissenschaftsverlage. Die Daten können verwendet werden, um bestehende Dienstleistungen zu verbessern. So können Wissenschaftlerinnen und Wissenschaftler beispielsweise aufgrund persönlicher Profile gezielt Vorschläge für die Rezeption und Hinweise auf Forschungsergebnisse aus ihrem Bereich automatisiert erhalten. Zudem können die Daten verwendet werden, um neue Dienstleistungen zu entwickeln. Neben der Bereitstellung und Verwaltung von Forschungsergebnissen in Form von wissenschaftlicher Literatur werden zunehmend Services in den Bereichen Forschungsdatenmanagement und Forschungssoftware angeboten.

Die verschiedenen Services können verknüpft sein, was ihre Nutzung für Forschende komfortabel macht. Wissenschaftlerinnen und Wissenschaftler können für zahlreiche oder sämtliche Aktivitäten im Forschungszyklus Dienste eines Anbieters nutzen, der zudem auch Angebote an ihre Einrichtungen macht (z. B. für Forschungsinformationssysteme). Zum Beispiel etabliert RELX, das Mutterunternehmen, zu dem Elsevier gehört, die Forschungsinformationssysteme-Software PURE in den Universitäten weltweit ausdrücklich mit dem Hinweis, Einblicke in den gesamten Forschungszyklus zu geben.[4] [5]

[4] Siehe die Angebotsbeschreibung unter: www.elsevier.com/solutions/pure.

[5] Elsevier erklärte am 22. September 2021: „PURE ist eine Software, die von institutionellen Kunden genutzt wird, um deren Daten zu prozessieren. Diese Daten gehören immer den Kunden und mit Beendigung des Vertrags erhalten die Kunden ihre Daten zurück (im Falle von Cloud-Hosting) oder die Daten verbleiben bei den Kunden (on-premise). Elsevier erwirbt keine Rechte an den Daten und nutzt die Daten in keinerlei Hinsicht. Wenn der Kunde das wünscht, hat Elsevier keinen Zugang zu der jeweiligen PURE-Installation, selbst wenn diese in der Cloud gehostet sein sollte. PURE und alle weiteren Elsevier-Produkte und Elsevier-Dienste sind DSGVO-konform. Die Software bietet alle Möglichkeiten für eine DSGVO-konforme Verarbeitung von personenbezogenen Daten, wenn sie richtig konfiguriert ist. Ab Sommer 2021 sind sowohl der Hosting-Partner Amazon als auch PURE nach ISO 27001 zertifiziert. Zudem schließt Elsevier mit allen Kunden Datenverarbeitungsvereinbarungen, die örtliche Anforderungen berücksichtigen. Eine Datenschutzvereinbarung existiert mit jeder Universität, der gegenüber Elsevier Dienstleistungen erbringt. Die Daten, die von den Kunden in PURE gehalten werden, sind isoliert von anderen PURE-Installationen und werden nicht von Elsevier genutzt, je sei denn die Universität aktiviert Datenaustauschfunktionen, die von der Universität kontrolliert werden. Wie bei allen Software-Produkten zeigen uns Prüfdaten an, ob die Software korrekt funktioniert."

Diese Entwicklung kann möglicherweise erheblich in die datenschutzrechtlich grundsätzlich gewährleistete Anonymität der Wissenschaftlerinnen und Wissenschaftler eingreifen und wissenschaftliche Institutionen zu Mitverantwortlichen für die Verletzung des Rechts auf informationelle Selbstbestimmung machen. Das Datentracking leistet potenziell auch dem Datenmissbrauch und der Wissenschaftsspionage Vorschub und kann zur persönlichen Diskriminierung von Wissenschaftlerinnen und Wissenschaftlern führen. Vor dem Hintergrund der aktuellen Rechtsprechung des BGH und des Schrems II-Urteils und des anstehenden Entwurfs für ein Plattformgesetz der EU (Digital Markets Act)[6] sollten Wissenschaftsorganisationen eine Position zu den Praktiken einnehmen.

Die kürzlich vorgelegte Datenstrategie der Bundesregierung geht nicht im Besonderen auf diese Situation ein, nennt aber grundsätzlich die Problematik: eine weitere Monopolisierung[7], Missbrauch von Marktmacht und Datenmissbrauch: „Bei der Nutzung von Daten ist nicht alles, was technisch möglich ist, auch ethisch vertretbar und politisch wünschenswert."[8]

Insgesamt entsteht für Forschende ein Spannungsverhältnis zwischen komfortablen gebündelten Services und der Kontrolle über ihre Daten. Vielfach sind sich Forschende der Bedeutung und Nutzung ihrer Daten als Wirtschaftsgut nicht bewusst. Zwischen den Polen des Komforts und der Kontrolle ist das Verhältnis auszuloten, das Wissenschaft und Verlage zueinander einnehmen. Dies kann jedoch nur vor dem Hintergrund klarer rechtlicher Regulierungen, mit hoher Transparenz und unter Mitbestimmung der Wissenschaft gelingen.

2. Transformation der Großverlage und ihr Verhältnis zur Wissenschaft

Verlage haben vor einiger Zeit damit begonnen, die Identifikation von Personen ermöglichenden Authentifizierungslösungen und User Trackings in ihre Angebote einzubauen. Dies ermöglicht ihnen, technisch proprietäre Services für den gesamten Forschungsprozess und die Analyse von Wissenschaftsdaten anzubieten. Ein Beispiel ist der 2020 in den Niederlanden geschlossene Vertrag mit Elsevier. Er sieht ausdrücklich Dienste, die als „Professional Services" bezeichnet werden, und die Sammlung von personenbezogenen Daten vor.[9] Manche

[6] So thematisiert der Digital Markets Act ausdrücklich die Zielsetzung, dass die gesammelten Daten nicht nur den Intermediären, sondern auch der Förderung des Wettbewerbs und des Gemeinwohls dienen sollen.

[7] Datenstrategie der Bundesregierung, Kabinettsfassung vom 27. Januar 2021, S. 21, Datenstrategie der Bundesregierung und die Ausschreibung des Bundesministeriums für Bildung und Forschung für Datentreuhandmodelle in den Bereichen Forschung und Wirtschaft vom 8. Januar 2021, Bekanntmachung - BMBF.

[8] Datenstrategie der Bundesregierung, Kabinettsfassung vom 27. Januar 2021, S. 7, Datenstrategie der Bundesregierung.

[9] Vertrag zwischen UKB und Elsevier 2020-2024: Signed UKB Elsevier SD 2020-2024 agreement.pdf (vsnu.nl). Entsprechende Passagen sind unter Schedule 5, aber auch Section 7.6. des Vertrags zu finden.

Verlage unterstützen die SeamlessAccess bzw. die GetFTR-Strategien,[10] die eine möglichst in sich geschlossene Informationsversorgung durch die großen Wissenschaftsanbieter vorsieht und eine einfache und einmalige Authentifizierung vorsehen.[11] GetFTR und SeamlessAccess bieten über ihre Webseiten Informationen dazu an, welche Daten erhoben werden und wie sie mit der Privatsphäre der Nutzenden umgehen.[12][13] Nach anfänglicher Kritik von Bibliotheksseite wurden Anpassungen vorgenommen.[14]

Aktuell haben deutsche Wissenschaftsorganisationen mit großen Wissenschaftsverlagen (Springer Nature[15] und Wiley[16]) DEAL-Verträge geschlossen, um Open Access und adäquate Preise für die Versorgung mit und Publikation von Forschungsergebnissen zu erzielen.

Wenn Verträge mit Verlagen geschlossen werden, sollte grundsätzlich ein Augenmerk auf die Vereinbarungen zu Data Privacy und zu den Zugangs- und Authentifizierungssystemen[17] gerichtet werden. Der grundsätzlich komfortabelste Zugang ist ein Zugang ohne Authentifizierung, das heißt im Open Access, wobei auch hier über Verlagsplattformen entsprechende Nutzungen getrackt werden könnten. Viele Einrichtungen sind zusätzlich zum Literaturzugang auch gebunden an Software z. B. von Elsevier.[18] Elsevier ist zudem Unterauftragnehmer der Europäischen Kommission, um in deren Auftrag Daten zu Open Science zu erheben (Open Science Monitor).[19]

Solche umfassenden Services bieten die Möglichkeit, Einblicke in möglichst alle Teile der Forschungsprozesse zu nehmen und weiter zu vermarkten, sodass Verlage bzw. Unternehmen zu denjenigen Instanzen werden könnten, welche die Akteure aus Wissenschaft und Politik,

---

[10] Siehe unter: www.getfulltextresearch.com und https://seamlessaccess.org

[11] Moore, S. A., „Individuation through infrastructure", in: *Journal of Documentation* 77(1) vom 28. Juli 2020, https://doi.org/10.1108/JD-06-2020-0090.

[12] GetFTR: GetFTR | Why GetFTR - GetFTR (getfulltextresearch.com) in FAQ no.7: https://www.getfulltextresearch.com/why-use-getftr/

[13] SeamlessAccess: Privacy and Trust - SeamlessAccess: https://seamlessaccess.org/about/trust/

[14] Hinchcliffe, L.J.: "Why are Librarians concerned about GetFTR?", in: *The Scholarly Kitchen* dated 10 November 2019, https://scholarlykitchen.sspnet.org/2019/12/10/why-are-librarians-concerned-about-getftr; Youngen, Ralph, Toler, Todd: "Lessons Learned: A Year with GetFTR", in: The Scholarly Kitchen dated 16 February 2021, https://scholarlykitchen.sspnet.org/2021/02/16/guest-post-lessons-learned-a-year-with-getftr/

[15] Kieselbach, S., *Projekt DEAL – Springer Nature Publish and Read Agreement*, 2020, https://doi.org/10.17617/2.3174351.

[16] Sander, F., Herrmann, G., Hippler, H., Meijer, G., & Schimmer, R., *Projekt DEAL – John Wiley & Son Publish and Read Agreement,* 2019, https://doi.org/10.17617/2.3027595.

[17] Stellungnahme des Deutschen Bibliotheksverbands „Empfehlungen zu Methoden zur Kontrolle des Zugriffs auf wissenschaftliche Informationsressourcen", www.bibliotheksverband.de/fileadmin/user_upload/DBV/positionen/2019_11_26_Rundgespaech_RA21_-_Stellungnahme_Empfehlungen_final.pdf.

[18] Vgl. die von Elsevier veröffentlichte Liste der Einrichtungen, die ihre Forschungsinformationssystem-Software Pure nutzen, www.elsevier.com/solutions/pure/clients.

[19] Siehe unter: Microsoft Word – Open Science Monitor Methodological Note April 2019.docx (europa.eu).

die Hochschulen und die Gesellschaft am umfassendsten und datenbasiert über wissenschaftliche Aktivitäten informieren können. Sie werden damit auch für die Steuerung von Wissenschaftseinrichtungen und Hochschulen unersetzbar. Von einem entstehenden „Superkontinent“[20] der wissenschaftlichen Informationsversorgung und der Versorgung mit Information über Wissenschaft ist bereits die Rede. Manche Daten über wissenschaftliche Aktivitäten können nützlich für die Wissenschaft selbst und für die komplexen Steuerungsprozesse moderner Wissenschaft sein. Eine gute Praxis wird z. B. dann erreicht, wenn die Regelungen der Datengewinnung, zur Nutzung der Daten und zur Weitergabe der Daten transparent und klar sind und die Daten auch zu nicht kommerziellen Zwecken den Akteuren in der wissenschaftlichen Intrastruktur selbst zur Verfügung stehen (z. B. bei CrossRef).

Die Folgen dieser „datengetriebenen Organisation von Wissenschaft“[21] und die Bedingungen zu ihrer Gewinnung sowie die Strukturen, die sie vorhalten, veräußern und verwerten, müssen letztlich in aller Konsequenz von der Wissenschaft selbst reflektiert und gestaltet werden. Die Wissenschaftsorganisationen sollten dafür einstehen, dass die Datensammlung und Datennutzung – wo sie nötig ist – nicht nur legal, sondern auch durch ethische Werte wie Transparenz und Nachvollziehbarkeit, Einwilligung bei vollumfänglicher Aufklärung über die Folgen und weitere Aspekte einer guten Datenpraxis geprägt sind und eine solche Datenpraxis zu einer Grundlage der Übereinkunft mit Anbietern machen.

2.1 Konsequenzen der Transformation der Verlage hin zu Data Analytics Businesses

Es besteht ein Risiko, dass die Wissensgesellschaft durch diese Verschiebung des kommerziellen Geschäftsmodells hin zur Datenanalytik privatisiert wird und letztlich nicht mehr die öffentliche Hand, sondern zunehmend private Unternehmen über das Wissen über Forschungsinhalte und -tendenzen, ihre Institutionen und Akteure verfügen. Wissenschaft als öffentliches Gut wird der Logik der Privatisierung von Infrastrukturen und ihren Folgen unterworfen.[22] Nicht nur Großverlage, sondern auch kleinere Anbieter von wissenschaftlichen Datenbanken sind Teil dieses Geschäfts. Verschiedene Untersuchungen und Initiativen wie schon der Aufruf „The Cost of Knowledge“ von 2012, aber auch Organisationen wie Science Europe[23] und Bib-

[20] Schonfeld, R. C.: „The Supercontinent of Scholarly Publishing?“, in: *The Scholarly Kitchen* vom 3. Mai 2018, https://scholarlykitchen.sspnet.org/2018/05/03/supercontinent-scholarly-publishing.

[21] Herb, U.: „Zw angsehen und Bastarde“, in*: Information. Wissenschaft & Praxis*, 69 (2-3), 2018, S. 87.

[22] Barlösius, E.: *Infrastrukturen als soziale Ordnungsdienste. Ein Beitrag zur Gesellschaftsdiagnose*. Frankfurt/M. 2019, Kapitel 6.4: „Infrastrukturierung der Forschung und infrastrukturierende Forschung“.

[23] „Science Europe calls for a clear exclusion of data users and usage for the purposes of research from the scope of the Digital Services Act, to ensure that unintended effects on research activity are avoided. A legislative act that aims to address the selling of illegal content on large commercial platforms could have side effects on sectors of public interest unless proper exceptions are introduced.“ Science Europe, The Digital Services Act

liotheksverbände haben wiederholt darauf aufmerksam gemacht, wie folgenreich dieser Informations- und Datenzuwachs bei privatwirtschaftlichen Firmen und die Konzentration des Wissens über die Wissenschaften nicht nur für Innovationen im Bereich der wissenschaftlichen Informationsversorgung sind.[24]

Die skizzierte Entwicklung hin zu einer privatwirtschaftlichen Wissensindustrie[25] steht im Gegensatz zur Wissenschaftsfreiheit, zum gesetzlich vorgegebenen Umgang mit personenbezogenen Daten und zum Wettbewerbsrecht. Im Einzelnen kann unreguliertes bzw. unerkanntes Datentracking

- eine Verletzung der Wissenschaftsfreiheit und der Freiheit von Forschung und Lehre bedeuten;
- eine Verletzung des Rechts auf den Schutz der eigenen Daten darstellen;
- eine potenzielle Gefährdung von Wissenschaftlerinnen und Wissenschaftlern darstellen, da die Daten auch ausländischen Regierungen und autoritären Regimes zugänglich werden können;
- einen Eingriff ins Wettbewerbsrecht darstellen, da neue Teilnehmer kaum eine Chance auf einen Markteintritt haben;
- eine Wertminderung öffentlicher Forschungsinvestitionen begünstigen, da im Rahmen von Wirtschaftsspionage wissenschaftliche Aktivitätsdaten von kommerziellen Forschungskonkurrenten erhoben oder ihnen gegen Bezahlung zugänglich gemacht werden können.

Wie kritisch diese Wissensindustrialisierung durch Tracking bereits ist, illustrieren erste Fälle des Datenhandels über Forschungsinteressen einzelner Wissenschaftlerinnen und Wissenschaftler.[26] LexisNexis, ein internationaler Anbieter von Informationslösungen und Tochterunternehmen der RELX Group, zu der auch Elsevier gehört, hat einen Vertrag unterzeichnet, durch den für 16,8 Millionen US-Dollar persönliche Daten an ICE, die amerikanische Behörde

---

Should Not Have Unintended Effects on Research, 2020, www.scienceeurope.org/media/4s3bnhbr/20200908_se_response_dsa_consultation_final.pdf.

[24] Z. B. Dobusch, L.: „Kein Open-Access-Deal, dafür Spyware gegen Schattenbibliotheken“, in: *netzpolitik.org* vom 26. November 2020, https://netzpolitik.org/2020/neues-vom-grossverlag-elsevier-kein-open-access-deal-dafuer-mit-spyware-gegen-schattenbibliotheken; die Stellungnahme des Deutschen Bibliotheksverbands „Empfehlungen zu Methoden zur Kontrolle des Zugriffs auf wissenschaftliche Informationsressourcen“, www.bibliotheksverband.de/fileadmin/user_upload/DBV/positionen/2019_11_26_Rundgespaech_RA21_-_Stellungnahme_Empfehlungen_final.pdf

[25] Burgelman, J.-C.: „Scholarly publishing needs regulation”, in: *Research Professional News* vom 28. Januar 2021, www.researchprofessionalnews.com/rr-news-europe-views-of-europe-2021-1-scholarly-publishing-needs-regulation.

[26] Jung, J.: „UCLA School of Law Holds Contract with Companies Selling Personal Data to ICE”, in: *The Daily Bruin* vom 17. Juli 2020, https://dailybruin.com/2020/07/17/ucla-school-of-law-holds-contracts-with-companies-selling-personal-data-to-ice.

für Immigration und Customs Enforcement, übergeben werden sollen.[27] Die Situation wird vielfach noch dadurch verkompliziert, dass Hochschulen und Bibliotheken auch ohne ihr Wissen zu Mitwirkenden in der Verletzung von Datenrecht, Wissenschaftsfreiheit und Wettbewerbsrecht werden können. Die Verhaltensdatenprofile deutscher Hochschulangehöriger können dabei in der gleichen Weise gehandelt und übermittelt werden, wie es im Schrems II-Urteil zur Aufhebung des Privacy Shields führte, also des Transfers persönlicher Daten in ein Drittland außerhalb der EU, da es sich um die gleichen Akteure handelt.[28]

Darüber hinaus könnten sich Risiken dadurch ergeben, dass die Großverlage sich mit einem zensierten Programm auf dem chinesischen Markt präsentieren. Aufgrund des Trackings könnten auch personalisierte Daten darüber entstehen, wer die zensierten Dokumente nutzt und weiterempfiehlt, ohne dass für die betroffenen Wissenschaftlerinnen und Wissenschaftler einschätzbar wäre, wem diese Trackingdaten zugänglich gemacht werden. In Reaktion auf mögliche Gesetzesänderungen hat z. B. Google kürzlich eine Änderung seiner Tracking-Policy angekündigt, die anonymer organisiert werden soll und auf deren Basis eher „Kohorten" statt einzelne Nutzerinnen und Nutzer identifiziert und adressiert werden sollen. [29]

3. Typen der Datengewinnung

Es sind drei unterschiedliche Typen, das heißt Verfahren, der Datengewinnung der von den Verlagen gesammelten und gespeicherten Daten denkbar (Third Party Data durch Microtargeting, Bidstream Data und Port Scanning sowie „Trojaner"), die im Folgenden beschrieben werden. Zudem gibt es unterschiedliche Werkzeuge: Tracker zu Seitenbesuchen, Audience-Tools zur Aggregierung verschiedener Datenquellen zu Profilen, Fingerprinter, die auch solche Nutzer identifizieren, die durch Browsereinstellungen eine Identifikation unterbinden wollen, und Werkzeuge zur Echtzeitversteigerung von Nutzerdaten machen das Portfolio der derzeit in den Wissenschaften zum Einsatz kommenden Werkzeuge aus. Die Werkzeuge zur Nachverfolgung stammen meistens von Drittanbietern der großen Internetfirmen, aber auch von spezialisierten Unternehmen wie der zu Oracle gehörenden Big-Data-Plattform BlueKai, die

[27] Biddle, S.: „LexisNexis to Provide Giant Database of Personal Data to ICE", in: *The Intercept* vom 2. April 2021, LexisNexis to Provide Giant Database of Personal Data to ICE (theintercept.com).

[28] Vgl. Bundesland Niedersachsen, *Das Schrems II-Urteil des Europäischen Gerichtshofs und seine Bedeutung für Datentransfer in Drittländer*, 2021, https://lfd.niedersachsen.de/startseite/themen/weitere_themen_von_a_z/internationaler_datenverkehr/das_schrems_ii_urteil_des_eugh_und_seine_bedeutung_fur_datentransfers_in_drittlander/das-schrems-ii-urteil-des-europaischen-gerichtshofs-und-seine-bedeutung-fur-datentransfers-in-drittlander-194085.html.

[29] Neue Spielregeln: Warum Google Cookie-Tracking abschafft (netzpolitik.org).

bereits Gegenstand von Sammelklagen wegen Missbrauchs personalisierter Daten ist.[30] Die Daten können, weil sie mit anderen Datenaggregatoren der Internetdienste schon institutionell verbunden sind, mit weiteren Daten aus anderen Lebensbereichen zu Profilen verdichtet werden.[31] Wie tief die Nachverfolgung reicht, legen die Verlage nicht offen, sodass hier derzeit nur auf verschiedene Tests verwiesen werden kann,[32] die etwa zeigen, dass, wer Artikel etwa in der Zeitschrift „Nature" aufruft, von mehr als 70 Instrumenten nachverfolgt wird.[33] Schließlich kommt noch hinzu, dass die genutzten Werkzeuge fehlerhaft sein und daher umso mehr nachteilige Effekte für einzelne Wissenschaftlerinnen und Wissenschaftler haben können.[34] Die oben genannten drei Hauptarten der Datengewinnung werden im Folgenden kurz skizziert. In der Summe ist davon auszugehen, dass die Instrumente zur Nachverfolgung der Wissenschaft beständig verfeinert und in ihrer Anwendung ausgebaut werden, da sie erhebliche Marktvorteile für die Anbieter und die Konzerne bedeuten.

3.1 Third Party Data durch Microtargeting

Microtargeting bezeichnet die zielgruppenspezifische Adressierung. Von den Verlagen werden Daten aus erster Hand und solche aus zweiter Hand genutzt. Daten aus erster Hand sind die direkten Nutzerspuren, Daten aus zweiter Hand sind dazu gekaufte Daten, die wiederum durch Dritte wie vor allem die großen Internetfirmen zu präzisen Datenprofilen verdichtet werden. Die Verlage haben eine große Vielfalt von diesen *Third Party Asset Sources* auf ihren Plattformen etabliert, seien es die gängigen Tracker von Google oder Facebook, seien es solche von Anbietern wie BlueKai und Krux Digital, seien es Browser-Fingerprinting-Tools wie Double Click oder datenaggregierende Audience Tools von Adobe, Neustar, Oracle, AddThis und andere. Der Javascript-Code der Third Parties hat Zugang zum Document Object Model der betreffenden Webseite und kann damit auslesen, mit welchem Text der Nutzer sich beschäftigt, zu welchem er als nächstes weiterbrowst und welche Suchworte er auf der Plattform eingibt. Da viele Anbieter teils die gleichen Third Parties einbinden, teils diese ihre Daten austauschen,

---

[30] Lomas, N.: „Oracle and Salesforce Hit with GDPR Class Action Lawsuits Over Cookie Tracking Consent", in: *TechCrunch* vom 14. August 2020, https://techcrunch.com/2020/08/14/oracle-and-salesforce-hit-with-gdpr-class-action-lawsuits-over-cookie-tracking-consent/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+Techcrunch+%28TechCrunch%29&guccounter=1.

[31] Vogel, C.: „Kennen Sie Google CASA?", in: *Medinfo. Informationen aus Medizin, Bibliothek und Fachpresse*, www.medinfo-agmb.de/archives/2020/07/08/6880.

[32] „Digital Library Federation, Endangering Data. Interview with Sarah Lamdan" s. u. www.diglib.org/endangering-data-interview-with-sarah-lamdan bzw. Lamdan, S.: „Social Media Privacy: A Rallying Cry to Librarians", in: *The Library Quarterly* 85 (3), 2015, S. 261-277, https://academicworks.cuny.edu/cl_pubs/52; Wolfie Christls Studien zu RELX und ThreatMetrix, https://twitter.com/wolfiechristl/status/1295655040741445632 und https://crackedlabs.org.

[33] Brembs, B.: https://twitter.com/brembs/status/1301897878387003398.

[34] Vgl. z. B. Lamdan, S.: „Librarianship at the Crossroad of ICE Surveillance", in: *In the library with the lead pipe* vom 13. November 2019 und Swauger, S.: https://twitter.com/SheaSwauger/status/1205587676172144641.

kann das Informationsverhalten von Hochschulangehörigen plattformübergreifend erhoben werden und im Fall von Google, Facebook oder Twitter mit dem bereits vorliegenden Wissen über ihr sonstiges Onlineleben verknüpft werden.[35] Im Fall von Anbietern wie Acxiom/Liveramp ist auch eine Synchronisation von Online- mit Offlineleben möglich, da dort Daten auch über Einkäufe, Führerscheine, TV-Konsum, Wählerverzeichnisse, Straffälligkeiten und anderes vorliegen.[36]

3.2 Bidstream Data und Port Scanning

Die Einbindung von Third Parties in Webseiten steht vielfach unter Kritik und wird teilweise von wichtigen Internetunternehmen und Einrichtungen nicht mehr unterstützt, sodass aktuell auch Alternativen wie das Harvesting von Bidstream Data (Real Time Bidding Data), also das im Hintergrund ablaufende Erheben von Daten über Ort, Geräte und genutzte Daten, genutzt werden. Bei den Echtzeitauktionen von Nutzerdaten können eine Vielzahl von Einzelinformationen wie Lokalisierungsdaten, IP-Nummer, Geräteinformationen und vieles mehr übertragen und mit einem Identifier verknüpft werden, um Personen auch ohne Setzen eines Cookies sicher identifizieren zu können.[37] Schon die Suche nach offenen Ports auf fremden Rechnern und/oder Netzwerken, um dort dann etwa Schad- oder Überwachungssoftware einzuschleusen, ist nach deutschem Verständnis am Rand der Legalität, da sie als Vorstufe von entsprechend sanktionierten Tatbeständen (§§ 202c, 303b StGB) gewertet werden kann. Dennoch wird sie vielfach eingesetzt, teils zur Betrugsprävention, teils als Trackinginstrument.

Ein öffentlich gewordenes Beispiel ist die Firma ThreatMetrix, die nach eigenen Angaben 4,5 Milliarden Geräte identifizieren kann und zu LexisNexis Risk Solutions/RELX gehört. ThreatMetrix ist etwa auf ScienceDirect implementiert, der Plattform, über welche Wissenschaftlerinnen und Wissenschaftler die Inhalte der Zeitschriften aus dem Elsevier-Verlag konsultieren. Die Verbindung des RELX-Konzerns zu verschiedenen staatlichen Stellen in den USA ist bereits Gegenstand von öffentlichen Petitionen in den USA.[38] Ob etwa Daten, die mit solchen

[35] Hanson, C.: *User Tracking on Academic Publisher Platforms*, 2019, www.codyh.com/writing/tracking.html.

[36] Vgl. die Grafik in Christl, W.: *Corporate Surveillance in Everyday Life*, S. 55, https://crackedlabs.org/dl/CrackedLabs_Christl_CorporateSurveillance.pdf.

[37] Vgl. Ryan, J.: *Briefing on adtech, RTB, and the GDPR at dmexco Brave Event*, Folie 45, www.slideshare.net/JohnnyRyan/briefing-on-adtech-rtb-and-the-gdpr-at-dmexco-brave-event.

[38] American Civil Liberties Union: *ACLU Calls On Tech Companies to End Their Alliance with ICE and CBP*, 2020, www.aclu.org/news/immigrants-rights/aclu-calls-on-tech-companies-to-end-their-alliance-with-ice-and-cbp.

Trackern erhoben werden, auch für andere Produkte der Risk-Solutions-Sparte[39], z. B. im Bereich der Analysen für Unternehmen und Behörden, genutzt werden, muss vermutet werden, solange die Verlage ihre Nachverfolgungspraktiken nicht aufdecken.[40]

3.3 „Trojaner“

Der Skalierung der Wissenschaftsnachverfolgung dienen „Trojaner“, die Bibliotheken im Zusammenhang mit Rabatten für andere Leistungen angeboten werden könnten. “Trojaner“ wären die in Bibliotheken zu installierende Zusatzsoftware, welche biometrische Daten wie Tippgeschwindigkeit oder Art der Mausbewegung sammelt, um auf diese Weise Nutzer trotz des Einsatzes von Proxy-Servern und VPN-Tunneln personalisieren zu können.[41] Organisationen und Anbieter könnten argumentieren, mit solcher Zusatzsoftware Nutzerinnen und Nutzer von „Schattenbibliotheken“[42] identifizieren und rechtlich verfolgen zu können.[43] Diese „Trojaner“ hebeln die Sicherheit von Hochschulnetzen aus und setzen die Hochschulen potenziell Angriffen aller Art aus. Die Nutzung solcher Software kann daher nicht empfohlen werden.[44]

---

[39] Risk & Business Analytics – RELX.

[40] Vgl. die Dokumentation von Wolfie Christl unter: https://twitter.com/wolfiechristl/status/1286341387718397952.

[41] Vgl. Mehta, G.: „Proposal to Install Spyware in Universities Libraries to Protect Copyrights Shocks Academics, in: *Coda* vom 13. November 2020, www.codastory.com/authoritarian-tech/spyware-in-libraries.

[42] Der Begriff ist eingebürgert und wird z. B. hier genutzt: Ball, R.: *Wissenschaftskommunikation im Wandel*, Springer, 2020, S. 127.

[43] In einer früheren Version des Informationspapiers (vom 20.05.2021) wurde PSI genannt. Diese Nennung wurde in der vorliegenden Version aufgrund einer Erklärung von PSI vom 09.06.2021 getilgt:

“1. *PSI arbeitet nicht mit SNSI in irgendeiner Art und Weise zusammen.*
2. *PSI trackt keine Nutzenden von Schattenbibliotheken.*
3. *PSI ist nicht an dem Einsatz von “Trojanern“ beteiligt und hat keine Kenntnisse darüber.*
4. *PSI hatte bislang keine Mitwirkung an der Strafverfolgung von „Nutzenden“ von Schattenbibliotheken und glaubt nicht, dass „Nutzende“ strafverfolgt worden sind.*

[44] In einer früheren Version des Informationspapiers wurde die Scholarly Networks Security Initiative (SNSI) in diesem Abschnitt genannt. Diese Erwähnung wurde in dieser Version getilgt aufgrund der Erklärung von SNSI vom 8. September 2021:

„SNSI ermutigt seine Kunden mit Nachdruck, wirkmächtige Schutzmechanismen für den Zugang zu Daten von Forschenden und Studierenden sowie die ihnen von seinen Mitgliedern angebotenen Inhalte aufrecht zu erhalten, aber SNSI hat und wird nicht:

- Die Nutzung von „Trojanern“ befürworten oder umsetzen (z.B. zur Sammlung von biometrischen Daten wie Tippgeschwindigkeit oder Art der Mausbewegung, um Nutzende trotz des Einsatzes von Proxy-Servern oder VPN-Tunneln zu personalisieren);
- Bibliotheken Anreize bieten oder für solche Anreize werben, um den Einsatz von „Trojanern“ durch Bibliotheken selbst zu befördern.“

4. Fazit

Diese Art der Nachverfolgung der Wissenschaft kann grundsätzlich der Wissenschaftsfreiheit und der informationellen Selbstbestimmung widersprechen. Sie kann Wissenschaftlerinnen und Wissenschaftler gefährden und die Wettbewerbsfreiheit im Bereich der Informationsversorgung behindern. Daher müssen sich Wissenschaftlerinnen und Wissenschaftler sowie wissenschaftliche Institutionen der Problematik bewusst werden und die rechtlichen wie technischen und ethischen Rahmenbedingungen ihrer Informationsversorgung klären – nicht zuletzt auch deshalb, um nicht unfreiwillig geltendes Recht zu verletzen, sondern ihre Wissenschaftlerinnen und Wissenschaftler zu informieren und zu schützen.

Der AWBI möchte mit diesem Informationspapier dazu anregen, einen breiten Diskurs in der Wissenschaft sowohl auf Ebene der wissenschaftlichen Entscheidungsträgerinnen und -träger als auch unter Wissenschaftlerinnen und Wissenschaftlern sowie in Einrichtungen der Informationsinfrastruktur zu führen, um die Praxis des Trackings, dessen Rechtmäßigkeit, Maßnahmen zur Einhaltung des Datenschutzes und Konsequenzen der Aggregation von Nutzungsdaten zu reflektieren und geeignete Maßnahmen zu ergreifen.

Die Erhebung von Daten über die Wissenschaft und über wissenschaftliche Aktivitäten kann sinnvoll sein, sofern sie transparenten und klaren Vorgaben folgt, Risiken für einzelne Wissenschaftlerinnen und Wissenschaftler minimiert werden sowie eine Nutzung solcher Daten durch, wenn nicht eine Hoheit über solche Daten bei wissenschaftlichen Organisationen gewährleistet werden kann.