

*Das **www.** (World Wide Web) hat in seiner Funktion als Produkt, Instrument, Gegenstand und Datenquelle das Wissenschaftssystem massiv verändert. Wenngleich Internetdaten, sogenannte digitale Spurdaten, für die sozial- und erziehungswissenschaftliche Forschung großes Potenzial haben, das von einer stetig wachsenden Zahl von Forschenden genutzt wird, haben sie auch gewisse Limitationen, auf die es zu reagieren gilt.*

Die Abkürzung **www.** steht für World Wide Web¹ und ist den meisten Menschen als üblicher Teil von Internetadressen geläufig (<https://www.>). **www.** und Wissenschaft sind auf verschiedenen Ebenen eng miteinander verwoben. Für nahezu alle Forschenden ist Wissenschaft ohne **www.** undenkbar geworden. Einerseits ist es ein mächtiges Werkzeug, welches Forschung erleichtert, beschleunigt oder erst ermöglicht. Andererseits sind **www.** und Internet selbst Ergebnisse wissenschaftlicher Forschung. Das **www.** wurde zwischen Ende der 1980er und Beginn der 1990er Jahre am CERN entwickelt; als Begründer gilt gemeinhin der britische Physiker und Informatiker Tim Berners-Lee. Auch der Vorläufer des Internets, das durch das US-Verteidigungsministerium geförderte Arpanet, wurde in den 1960er und 1970er Jahren im Wissenschaftskontext entwickelt und diente ursprünglich der Vernetzung von Computern an Universitäten und anderen Forschungseinrichtungen. Hieran wird deutlich, dass die Genese, Entwicklung sowie die originäre Nutzung des **www.** in der Wissenschaft verortet sind. Diese strukturelle Kopplung ist einer der Gründe, warum das **www.** insbesondere die Wissenschaft in den letzten Jahrzehnten nicht nur geprägt, sondern grundlegend verändert hat.

Das **www.** und Services, die darauf basieren und zugreifen (etwa Mail, Suchmaschinen oder Online-Datenbanken), werden von Forschenden für alle Bereiche und Phasen ihrer wissenschaftlichen Tätigkeit genutzt: Für das Suchen nach Forschungsliteratur, den Zugriff auf Literatur oder andere relevante Informationen, die Kommunikation mit anderen Forschenden, die Zusammenarbeit an Dokumenten oder anderen Erzeugnissen, wie etwa Software, oder für die Veröffentlichung und Verbreitung von Ergebnissen und Erkenntnissen. Zur letzten Kategorie gehören auch die Bekanntgabe und Kommunikation von Forschungsergebnissen über Social Media, Blogs und andere Plattformen. Einer der eindeutigen Indikatoren für die enorme Bedeutung des **www.** für die Wissenschaft ist, dass viele wissenschaftliche Journale mittlerweile ausschließlich online erscheinen und es kaum noch Journale gibt, die ausschließlich in gedruckter Form publiziert werden. Die Digitali-

sierung wissenschaftlicher Erzeugnisse (Daten, Publikationen et cetera) erleichtert ihre Verbreitung und damit potenziell auch den Zugriff sowohl für Forschende als auch für die Öffentlichkeit. Die Verbreitung von Onlineservices und **www.** sowie ihre Nutzung durch die Wissenschaft haben in wesentlichem Maße dazu beigetragen, dass Reformbewegungen im Hinblick auf den Zugang zu wissenschaftlichen Erzeugnissen wie Daten (Open Science) und Publikationen (Open Access) entstehen konnten.

Neben der Rolle als Forschungswerkzeug nimmt das Internet für die wissenschaftliche Arbeit mindestens zwei weitere wichtige Funktionen ein, die mit der fortschreitenden Digitalisierung aller Lebensbereiche an Bedeutung gewonnen haben: (1) als Erkenntnisgegenstand und (2) als Datenquelle.

Das Internet als Erkenntnisgegenstand

Eine wichtige Funktion des Internets für die Wissenschaft ist diejenige als Erkenntnisgegenstand. In den Sozial-, Erziehungs- und Geisteswissenschaften werden vornehmlich die gesellschaftlichen sowie individuellen Nutzen, Risiken und Auswirkungen des Internets untersucht (Salganik 2018). Das Internet ist als Erkenntnisgegenstand von solcher Relevanz für eine Vielzahl von Wissenschaftsdisziplinen, dass sich eigene Felder und Organisationen für seine inter- und transdisziplinäre Erforschung etabliert haben.² Angesichts der Omnipräsenz des Internets in wissenschaftlichen und außerwissenschaftlichen Bereichen ist die Spannweite der Fragen innerhalb der Forschung, die sich mit dem **www.** befasst, nicht verwunderlich. In den Erziehungs- und Sozialwissenschaften liegt der Fokus in weiten Teilen auf Fragen der Nutzung (*Was machen Menschen im Internet?*) und Wirkung (*Wie beeinflusst das Internet die Nutzer*innen?*). Dabei werden die Forschungsfragen nicht immer nur auf der Individual- oder Mikroebene gestellt, sondern können auch eine Meso- oder Makroperspektive haben. Beispiele für diese Perspektiven sind Untersuchungen der Auswirkungen des Internets auf die Institution Schule oder das gesamte Wissenschaftssystem. Ebenso vielfältig wie die Fragestellungen sind die Methoden, mit denen Nutzung und Wirkung des Internets beziehungsweise seine Inhalte und Anwendungen in den sozial- und erziehungswissenschaftlichen Disziplinen untersucht werden. Neben diskursanalytischen Ansätzen kommt in der empirischen Erforschung dieses Bereichs die gesamte Bandbreite qualitativer und quantitativer Methoden zum Einsatz. Entsprechend heterogen sind auch die verwendeten Forschungsdaten. So werden Interview-, Befragungs-, und Experimentaldaten ebenso genutzt wie inhaltsanalytische Daten.

Das Internet als Datenquelle

Zusätzlich zur Rolle als Erkenntnisgegenstand ist eine weitere Funktion des Internets die der Datenquelle. Sie hat für die Wissenschaft und speziell die sozial- und geisteswissenschaftliche Forschung in jüngster Zeit enorm an Bedeutung gewonnen. Diese Entwicklung bezieht sich einerseits auf die zunehmende Digitalisierung von Quellen (zum Beispiel historische Texte oder Nachrichteninhalte), steigende Verfügbarkeit von digitalen Forschungsdaten und die fortlaufende Verbreitung der Ideale einer Open Science sowie die damit verbundenen Veränderungen von Erwartungen und Anforderungen (etwa seitens der Journale und Forschungsförderer). Andererseits generiert die Internetnutzung sowie von Anwendungen und Services, die darauf zugreifen, eine große Menge an Daten, welche von Anbieter*innen unter anderem genutzt werden, um Angebote zu verbessern und mehr über Nutzer*innen zu erfahren.³ Diese sogenannten digitalen Spurdaten (*digital trace data*) oder digitalen Verhaltensdaten (*digital behavioral data*) sind auch für die wissenschaftliche Forschung von großem Interesse und werden von Forschenden zusehends genutzt, um sowohl neue Fragen als auch neue Antworten auf bekannte Fragen zu erarbeiten. Der Gebrauch digitaler Spurdaten ist kennzeichnend für die noch jungen, aber rasant wachsenden Felder der Digital Humanities und Computational Social Science (Heiberger/Riebling 2016; Hox 2017). Howison, Wiggins und Crowston (2011: 769) definieren digitale Spurdaten als „records of activity (trace data) undertaken through an online information system (thus, digital)“. In ähnlicher Weise beschreibt Bail (2020: o.S.) sie als „large digital datasets that describe human behavior“.

Aufgrund ihres Umfangs werden digitale Spurdaten häufig der Kategorie Big Data zugeordnet. Wenngleich der Begriff *Big Data* oft unterdefiniert ist,⁴ treffen die sogenannten *Three Vs of Big Data* auf digitale Spurdaten prinzipiell zu: Sie entstehen im großen Umfang (Volume) und mit hoher Geschwindigkeit (Velocity), wobei sie sehr unterschiedliche Formate haben können (Variety). In ihrer Entstehung unterscheiden sich digitale Spurdaten zudem von anderen Arten von Daten, die in der empirischen sozial- und erziehungswissenschaftlichen Forschung gemeinhin zum Einsatz kommen. Anders als beispielsweise Interview- oder Befragungsdaten werden digitale Spurdaten nicht originär für und durch die Forschung selbst produziert, sodass die Entstehung der Daten und ihre wissenschaftliche Nutzung voneinander getrennt sind. Aus diesem Grund werden digitale Spurdaten häufig auch als *found data* (in Abgrenzung zu *designed data*) oder, in Anlehnung

an die bildenden Künste, als *ready-made data* (in Abgrenzung zu *custom-made data*) bezeichnet (ebd.). Neben dem zugrundeliegenden datengenerierenden Prozess ist Vielfalt ein weiteres wesentliches Merkmal digitaler Verhaltensdaten. Digitale Verhaltensdaten können durch die Nutzung verschiedener Plattformen und Anwendungen erzeugt werden und in zahlreichen Formaten vorliegen. Ein prominentes Beispiel sind Daten von Social-Media-Plattformen wie Twitter, Facebook, Instagram oder YouTube, die in der aktuellen Forschung, speziell der Computational Social Science, häufig verwendet werden. Je nach Plattform und Art der Datensammlung kann es sich beispielsweise um Textdaten, Netzwerkdaten oder Bilder handeln. Im Hinblick auf ihren Entstehungskontext unterscheiden sich digitale Spurdaten zudem dahingehend, ob sie von Nutzer*innen bewusst oder unbewusst erzeugt wurden. Hox (2017) unterscheidet zwischen *intentional digital traces* (etwa Posts, Kommentare) und *unintentional digital traces* (etwa automatisch erfasste Standort- und Gerätedaten). In ähnlicher Weise differenziert Menchen-Trevino (2013) *participation traces* (etwa Posts, Kommentare) und *transactional data* (etwa Daten zu Logins). Bezüglich der Sammlung digitaler Verhaltensdaten unterscheidet Menchen-Trevino zudem zwischen *horizontal trace data* und *vertical trace data*: Während erstere in die Breite gehen, wie im Falle einer Sammlung aller Tweets zu einem Hashtag, zielen zweite auf die Tiefe ab, insofern sie eine größere Menge unterschiedlicher Daten für dieselben Beobachtungseinheiten erfassen. Ein Beispiel für *vertical trace data* wäre die Sammlung umfangreicher Daten, etwa von unterschiedlichen Social-Media-Plattformen und aus Interviews, für eine begrenzte Gruppe von Nutzer*innen. An dieser Unterscheidung wird deutlich, dass die Weise, auf welche die Daten gesammelt werden, maßgeblich bestimmt, was mit ihnen möglich ist.

Potenziale und Limitationen digitaler Spurdaten

Für die wachsende Popularität digitaler Spurdaten in der sozial- und erziehungswissenschaftlichen Forschung und anderen Fachbereichen gibt es mehrere Gründe. Zum einen werden diese Daten in enormem Umfang und mit hoher Geschwindigkeit produziert; zum anderen ist der Umgang mit ihnen durch das rasante Wachstum von Rechenleistung und die (Weiter-)Entwicklung von Analyseverfahren für Forschende deutlich einfacher geworden. Die Nutzung solcher etwa aus den Bereichen Machine Learning und Natural Language Processing stammenden Verfahren ist laut Hox (2017) ein wesentliches Merkmal

der Computational Social Science. Aufgrund dieser Entwicklungen haben Forschende „the capacity to collect and analyze data with an unprecedented breadth and depth and scale“ (Lazer et al. 2009: 722). Auch im direkten Vergleich mit anderen Datentypen, die in den Sozial- und Erziehungswissenschaften verbreitet sind, bieten digitale Spurdaten einige Vorteile. Im Gegensatz zu Daten aus Selbstauskünften sind sie weniger (direkt) von sozialer Erwünschtheit beeinflusst. Während bei Selbstauskünften das Verhalten retrospektiv erfasst wird, erfolgt die Erfassung digitaler Verhaltensdaten zudem instantan, sodass für digitale Spurdaten nicht das Problem der Verzerrung durch Erinnerungsfehler besteht. Mehrere Studien konnten zeigen, dass Selbstauskünfte zur Mediennutzung aus diesem Grund wenig verlässlich sind (Araujo/Wonneberger/Neijens/de Vreese 2017; Scharkow 2016). Insbesondere bei spezifischen und seltenen Verhaltensweisen oder solchen, die lange zurückliegen, ist die Verzerrung durch Erinnerungsfehler ein wesentliches Problem.

Wenngleich digitale Spurdaten für die sozial- und erziehungswissenschaftliche Forschung großes Potenzial haben, sind auch sie mit spezifischen Limitationen verbunden. So fehlen in digitalen Spurdaten oftmals detaillierte Informationen zu den individuellen Nutzer*innen. Während viele Plattformen Informationen zu grundlegenden demografischen Charakteristika liefern können (Alter, Geschlecht, Wohnort), beinhalten digitale Spurdaten dennoch keine oder sehr begrenzt direkte Informationen zu Meinungen oder Einstellungen, worin sie sich von Interview- oder Befragungsdaten unterscheiden. Darüber hinaus sind auch für die Sozial- und Erziehungswissenschaften relevante abhängige Variablen wie Wahlverhalten, Offline-Aktivitäten oder Indikatoren zur schulischen Leistung häufig nicht in digitalen Spurdaten enthalten. Eine weitere Limitation sind mögliche Verzerrungen (*biases*) in digitalen Spurdaten (Jürgens/Stark/Magin 2020; Sen/Flöck/Weller/Weiß/Wagner 2021). Diese können durch die Stichprobenziehung (*sampling*) sowie die Auswahl von Datentypen und Operationalisierung von Messungen entstehen (siehe auch *, Scharlau/Jenert 2023). Abhängig von Art und Ausmaß der Verzerrungen sind daher Erkenntnisse, die mit Daten von Nutzer*innen bestimmter Plattformen gewonnen werden, möglicherweise nicht auf andere Populationen übertragbar. Um dies zu vermeiden, sollten (mögliche) Verzerrungen bei der Planung der Datenerhebung und -auswertung berücksichtigt werden. Überdies stellen sich epistemologische Fragen bezüglich der Analyse und Interpretation bestimmter Variablen. So müssen etwa Retweets oder Likes bei Twitter angesichts ihrer unterschiedlichen Nutzung nicht unbedingt Zustimmung

zu deren Inhalten bedeuten. Es geht in diesem Kontext entsprechend um die Frage, welche Informationen aus den Daten gewonnen werden, sowie darum, wie und wozu diese genutzt werden können.

Eine Möglichkeit, den Limitationen digitaler Spurdaten zu begegnen, ist ihre Verknüpfung mit anderen Datentypen. So ermöglichen beispielsweise Kombinationen mit Befragungsdaten umfassendere Einblicke in Eigenschaften und Einstellungen von Individuen (Stier/Breuer/Siegers/Thorson 2020). Auch können qualitative Interviewdaten wichtige Kontextinformationen zu Verständnisprozessen und Intentionen bei bestimmten Nutzungsweisen von Plattformen und Services im Internet liefern. Ob und welche Verknüpfungen möglich sind, hängt wesentlich davon ab, wie die Spurdaten gesammelt wurden und wie viel Einblick und Kontrolle die Forschenden im Prozess der Datengenerierung haben.

Zugangswege zu digitalen Spurdaten

Entsprechend der Vielfalt der Quellen und Typen digitaler Spurdaten gibt es für Forschende verschiedene Möglichkeiten des Zugangs. Breuer, Bishop und Kinder-Kurlanda (2020) unterscheiden zunächst drei Arten: Forschende können (1) Daten selbst erheben, (2) direkt mit Anbieter*innen von Plattformen/Services kooperieren oder (3) Daten von Drittanbieter*innen erwerben. Bei der eigenständigen Erhebung kann zudem danach unterschieden werden, ob die Daten manuell über Copy/Paste oder Screenshots, mittels Web Scraping oder über die von vielen Plattformen angebotenen Application Programming Interfaces (APIs) gesammelt werden. Alle Zugangsweisen unterscheiden sich in verschiedenen Dimensionen und haben eigene Vor- und Nachteile. Die wesentlichen Unterschiede aus Sicht der Forschenden beziehen sich auf die für die jeweilige Zugangsmethode benötigten Ressourcen (Wissen, Fähigkeiten, Zeit, Kosten) sowie den Typ und Umfang der Daten, die gewonnen werden.

An den Beispielen der Zugangswege wird deutlich, dass letztlich die Plattformen – meist kommerzielle Anbieter*innen – den Zugang zu den Daten kontrollieren. Angesichts dieser Einschränkung ergibt sich für Forschende eine paradoxe Situation: „Social scientists have access to more data than they ever had before to study human society, but a far smaller proportion than at any time in history“ (King/Persily 2018: 3). Die Auswirkungen der Abhängigkeiten von den Entscheidungen kommerzieller Plattform- und Service-Anbieter*innen im Internet zeigen sich beispielsweise an der drastischen Reduktion des Datenzugangs über die Facebook-API im Zuge des Cambridge-Analytica-Skandals. Diese hat die

Forschung mit über die API gesammelten Daten für Facebook weitgehend unmöglich gemacht. In Anbetracht dieser Entwicklungen sowie des grundsätzlichen Risikos, dass Unternehmen, die Plattformen und Services betreiben, den Datenzugang über APIs jederzeit reduzieren und beenden können, sieht Free-lon (2018) ein „post-API age“ für Computational Research anbrechen. Die Verfügbarkeit von Daten über APIs beeinflusst nicht zuletzt auch, welche Plattformen eingehender erforscht werden können. So ist die relative Freizügigkeit im Zugang zu Daten über die API einer der Gründe dafür, dass Twitter so umfassend beforscht wird. Ein weiteres Problem, dass sich durch die Anforderungen für den Zugriff auf digitale Verhaltensdaten ergibt, ist die Entstehung von Disparitäten im Datenzugang. boyd und Crawford (2012: 674) sehen die Gefahr einer Aufteilung in „Big Data rich“ und „Big Data poor“, welche zu Ungleichheiten in den Forschungsmöglichkeiten führt. So besteht bei den aktuell verbreiteten Zugangswegen das Risiko, dass nur Forschende beziehungsweise Institutionen mit den entsprechenden Ressourcen (Fähigkeiten, Technologie, Geld) oder Kontakten zu Plattformen Zugang zu den für sie relevanten digitalen Spurdaten erhalten.⁵

Aufgrund der skizzierten Einschränkungen und Risiken werden von Forschenden alternative Modelle des Datenzugangs vorgeschlagen und diskutiert. Einige Vorschläge zielen auf neue Kooperationsmodelle zwischen Forschenden und Firmen (King/Persily 2019; Puschmann 2019). Eine weitere Möglichkeit des Datenzugangs besteht in dem, was Halavais (2019: 8) in seinem „proposal for ethical distributed research“ als „partnering with users to collect big data“ beschreibt. In einem solchen Szenario kooperieren Forschende nicht mit Betreiber*innen von Plattformen und Services, sondern direkt mit Nutzer*innen. An anderen Stellen wird ein solcher Ansatz auch als Datenspende (*data donation*) bezeichnet (Boeschoten/Ausloos/Moeller/Araujo/Oberski 2020). Die meisten Plattformen bieten Nutzer*innen die Option, eigene Daten zu exportieren.⁶ Weil die Nutzer*innen diese selbstständig mit Forschenden teilen, erzeugt dieser Ansatz mehr Transparenz für sie. Er umgeht zudem die Problematik der Einschränkungen des Datenzugriffs und der Datennutzung durch technische Maßnahmen (etwa Einschränkungen der Zahl und/oder des Umfangs der Zugriffe über APIs) oder Nutzungsbedingungen (Terms of Service). Selbstverständlich bringt auch die Kooperation mit Nutzer*innen über eine Datenspende eigene Herausforderungen und Limitationen mit sich. Diese betreffen etwa die Rekrutierung der Teilnehmer*innen, die technische Umsetzung und den Datenschutz.

Eine weitere Möglichkeit des Datenzugangs ist die Nachnutzung (Sekundäranalyse) von bestehenden Daten aus abgeschlossenen oder fortlaufenden Sammlungen digitaler Verhaltensdaten. Aber auch diese Daten müssen erst gesammelt werden. Zu bedenken ist ferner, dass die Veröffentlichung und Nachnutzung von Daten durch Datenschutzvorgaben oder Nutzungsbedingungen von Plattformen sowie ihrer APIs eingeschränkt sein kann. An diesem Punkt wird deutlich, dass sich für digitale Spurdaten in allen Phasen des klassischen Forschungsdatenzyklus – von der Studienplanung über die Datenerhebung und -auswertung bis hin zur Publikation – spezifische Herausforderungen ergeben, die unterschiedliche forschungspraktische, ethische und rechtliche Fragen betreffen. Trotz dieser Herausforderungen bietet die Nutzung digitaler Verhaltensdaten für die sozial- und erziehungswissenschaftliche Forschung (wie auch andere Disziplinen) enorme Chancen, ihr Portfolio an Methoden und Fragestellungen zu erweitern. Die Bedeutung des Internets als Datenquelle wird weiterhin wachsen, wodurch auch seine Funktionen als Werkzeug sowie Erkenntnisgegenstand nochmals an Bedeutung gewinnen.

Fazit

Das Internet nimmt im Verhältnis zu Wissenschaft und Forschung verschiedene Rollen ein, deren Eigenschaften und Bedeutung sich über die Zeit verändert haben. Das Internet ist zugleich Produkt, Werkzeug und Erkenntnisgegenstand der Wissenschaft. Aufgrund der fortschreitenden Digitalisierung sowie der enorm gestiegenen Produktion und Relevanz von Daten ist es in jüngster Zeit darüber hinaus zu einer wichtigen Quelle für Forschungsdaten geworden. Digitale Spurdaten, welche durch die Nutzung digitaler Plattformen und Services entstehen, werden von Forschenden aus verschiedenen Fachbereichen und speziell in den sozial- und erziehungswissenschaftlichen Disziplinen vermehrt genutzt. Im Vergleich zu anderen etablierten Methoden der Datengewinnung bieten sie eine Reihe von Vorteilen, haben gleichsam jedoch spezifische Limitationen, weshalb eine Verknüpfung verschiedener Datentypen und Methoden gewinnbringend sein kann. Wie bei anderen Datentypen erzeugt die Analyse und Interpretation digitaler Spurdaten Informationen, die wiederum zur Produktion von Wissen genutzt werden können. Aufgrund ihrer Menge und Vielfalt, der Geschwindigkeit ihrer Entstehung sowie ihrer (potenziellen) Verfügbarkeit ist es möglich, die Informationsgewinnung und Wissensproduktion mithilfe von digitalen Spurdaten zu erweitern und beschleunigen. Hierfür ist es wichtig, dass möglichst alle Forschenden glei-

chen und möglichst freien⁷ Zugang zu diesen Daten haben. Hier gilt es, Disparitäten im Datenzugang zu reduzieren und Modelle des Zugangs zu entwickeln, welche Wissenschaftler*innen aus allen Regionen, Disziplinen und Institutionen den Zugriff auf diejenigen digitalen Verhaltensdaten ermöglichen, die sie für ihre Forschung benötigen. Solche Entwicklungen in Richtung einer offenen Forschungskultur im Sinne von Open Science können dazu beitragen, dass die Bedeutung des Internets für die Wissenschaft als Quelle für Daten und Informationen sowie als Werkzeug zur Produktion und Dissemination von Wissen weiter wächst – und somit Forschung ermöglicht, erweitert und verbessert. Mit dem Voranschreiten der Digitalisierung ist zu erwarten, dass die Menge digitaler Spurdaten weiterhin steigen wird. Die zunehmende Bedeutung von Daten für viele Lebensbereiche, die oftmals als Datafizierung (*datafication*) beschrieben wird, verändert auch die Wissenschaft. Dies ist einerseits eine Chance, zugleich jedoch eine Herausforderung, der die Wissenschaft begegnen muss, um das produktive Potenzial des Internets für die Forschung auszubauen und weiterzuentwickeln.

Anmerkungen

- Alltagssprachlich werden die beiden Begriffe *World Wide Web* (www.) und *Internet* oftmals synonym verwendet. Technisch gesehen ist das www. jedoch nur ein Teil beziehungsweise eine Anwendungsmöglichkeit des Internets; die bekannteste andere Internetanwendung ist E-Mail. Sofern diese Unterscheidung nicht explizit getroffen wird, werden im weiteren Verlauf des Beitrags die Begriffe im Sinne der besseren Lesbarkeit synonym verwendet.
- Gängige Bezeichnungen für diese Forschungsfelder sind beispielsweise *Web Science* oder *Internet Research* und es sind in den letzten Jahrzehnten zahlreiche Journale, Konferenzen und Fachgesellschaften entstanden, die sich diesen Feldern zuordnen und teilweise auch deren Bezeichnungen im Titel tragen, etwa die Association of Internet Researchers (AoIR) oder die ACM Web Science Conference (WebSci).
- In der Regel ist das Ziel kommerzieller Anbieter*innen in beiden Fällen die Maximierung von Profit. Auch wenn dies zweifelsohne ein wichtiges Thema ist, wird die ethische Bewertung dieser Datensammelpraktiken seitens der Unternehmen hier nicht weiter diskutiert.
- Aufgrund der Komplexität und häufig uneinheitlichen Definition des Begriffs gibt es Forschung, die zum Ziel hat, die (für die Wissenschaft) relevanten Merkmale und Besonderheiten von Big Data zu erfassen. Das interdisziplinäre Projekt ABIDA – Assessing Big Data (siehe <https://www.abida.de/>) ist ein gutes Beispiel hierfür.
- Während die Ungleichheiten im Hinblick auf Kontakte und finanzielle Ressourcen deutlich schwerer zu beseitigen sind, ist im Hinblick auf die Fähigkeiten, die benötigt werden, um digitale Verhaltensdaten zu erheben und analysieren, zu erwarten, dass sich Unterschiede zwischen Forschenden und Institutionen verringern, da es eine stetig steigende Zahl von Tutorials, Workshops und anderen Fortbildungsangeboten für diesen Bereich gibt und Digital Methods sowie Computational Methods zusehends Eingang in die Curricula von Studiengängen finden.
- In den meisten Fällen wurden diese Optionen als Reaktion auf die Einführung der Datenschutzgrundverordnung (DSGVO) in der EU eingeführt oder erweitert.
- Selbstverständlich ist ein völlig freier Zugang zu digitalen Spurdaten nicht möglich, da rechtliche Rahmenbedingungen (insbesondere im Hinblick auf den Datenschutz) auch in der wissenschaftlichen Forschung beachtet werden müssen, obschon es für die akademische Forschung teilweise gesonderte Regelungen gibt (unter anderem innerhalb der DSGVO).

Referenzen

- Araujo, Theo/Wonneberger, Anke/Neijens, Peter/de Vreese, Claes (2017). How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-reported Measures of Internet Use. *Communication Methods and Measures*, 11(3), 173–190.
- Bail, Chris (2020). What Is Digital Trace Data? Video, YouTube. URL: [youtube.com/watch?v=uuSWQN7uYhk](https://www.youtube.com/watch?v=uuSWQN7uYhk) [1.6.2021]
- Boeschoten, Laura/Ausloos, Jef/Moeller, Judith/Araujo, Theo/Oberski, Daniel L. (2020). Digital Trace Data Collection Through Data Donation. *ArXiv:2011.09851 [Cs, Stat]*.
- boyd, danah/Crawford, Kate (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society*, 15(5), 662–679.

- Breuer, Johannes/Bishop, Libby/Kinder-Kurlanda, Katharina (2020). The Practical and Ethical Challenges in Acquiring and Sharing Digital Trace Data. Negotiating Public-Private Partnerships. *New Media & Society*, 22(11), 2058–2080.
- Freelon, Deen (2018). Computational Research in the Post-API Age. *Political Communication*, 35(4), 665–668.
- Halavais, Alexander (2019). Overcoming Terms of Service: A Proposal for Ethical Distributed Research. *Information, Communication & Society*, 22(11), 1567–1581.
- Heiberger, Raphael H./Riebling, Jan R. (2016). Installing Computational Social Science: Facing the Challenges of New Information and Communication Technologies in Social Science. *Methodological Innovations*, 9, 1–11.
- Howison, James/Wiggins, Andrea/Crowston, Kevin (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, 12(12), 767–797.
- Hox, Joop (2017). Computational Social Science Methodology, Anyone? *Methodology*, 13 (Supplement 1), 3–12.
- Jürgens, Pascal/Stark, Birgit/Magin, Melanie (2020). Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data. *Social Science Computer Review*, 38(5), 600–615.
- King, Gary/Persily, Nathaniel (2019). A New Model for Industry-Academic Partnerships. *PS: Political Science & Politics*, 53(4), 703–709.
- Lazer, David/Pentland, Alex/Adamic, Lada/Aral, Sinan/Barabasi, Albert Laszlo/Brewer, Devon/Christakis, Nicholas/Contractor, Noshir/Fowler, James/Gutmann, Myron/Jebara, Tony/King, Gary/Macy, Michael/Roy, Deb/Van Alstyne, Marshall (2009). Life in the Network: The Coming Age of Computational Social Science. *Science*, 323(5915), 721–723.
- Menehen-Trevino, Ericka (2013). Collecting Vertical Trace Data: Big Possibilities and Big Challenges for Multi-Method Research. *Policy & Internet*, 5(3), 328–339.
- Puschmann, Cornelius (2019). An End to the Wild West of Social Media Research: A Response to Axel Bruns. *Information, Communication & Society*, 22(11), 1582–1589.
- Salganik, Matthew J. (2018). *Bit by Bit: Social Research in the Digital Age*. Princeton: University Press.
- Scharlau, Ingrid/Jenert, Tobias (2023). *. In Sandra Hofhues/Kons-tanze Schütze (Hg.), *Doing Research*. Bielefeld: Transcript, 44–51.
- Scharkow, Michael (2016). The Accuracy of Self-Reported Internet Use – A Validation Study Using Client Log Data. *Communication Methods and Measures*, 10(1), 13–27.
- Sen, Indira/Flöck, Fabian/Weller, Katrin/Weiß, Bernd/Wagner, Claudia (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly*.
- Stier, Sebastian/Breuer, Johannes/Siegers, Pascal/Thorson, Kjerstin (2020). Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Social Science Computer Review*, 38(5), 503–516.

