Big Data & New Data: Ein Ausblick auf die Herausforderungen im Umgang mit Social-Media-Inhalten als neue Art von Forschungsdaten

Weller, Katrin

Auszug aus dem Buch:

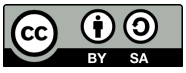
Uwe Jensen
Sebastian Netscher
Katrin Weller (Hrsg.)

Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten

Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten

Verlag Barbara Budrich
Opladen • Berlin • Toronto 2019

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
http://dnb.d-nb.de abrufbar.



Dieses Buch steht im Open-Access-Bereich der Verlagsseite zum kostenlosen Download bereit (https://doi.org/10.3224/84742233).
Eine kostenpflichtige Druckversion (Print on Demand) kann über den Verlag bezogen werden. Die Seitenzahlen in der Druck- und Onlineversion sind identisch.

ISBN 978-3-8474-2233-4 (Paperback)
eISBN 978-3-8474-1260-1 (eBook)
DOI 10.3224/84742233

Umschlaggestaltung: Bettina Lehfeldt, Kleinmachnow – www.lehfeldtgraphic.de
Lektorat: Nadine Jenke, Potsdam
Satz: Anja Borkam, Jena – kontakt@lektorat-borkam.de
Titelbildnachweis: Foto: Florian Losch
Druck: paper & tinta, Warschau
Printed in Europe

11. Big Data & New Data: Ein Ausblick auf die Herausforderungen im Umgang mit Social-Media-Inhalten als neue Art von Forschungsdaten

Katrin Weller

Seit ungefähr einem Jahrzehnt wird (auch) in wissenschaftlichen Kontexten der Nutzen von neuartigen, großen Datenbeständen für das bessere Verständnis zahlreicher Lebensbereiche erprobt. Viele dieser neuartigen Daten stammen aus Internetplattformen. In den Fokus der Wissenschaftler/innen rücken beispielsweise Suchmaschinen (Choi/Varian 2012), Kommentarbereiche von Zeitungen (Ruiz et al. 2011) sowie Social-Media-Plattformen (z.B. Facebook, LinkedIn, reddit, Twitter, Pinterest oder Tumblr). Social-Media-Plattformen können aus verschiedenen Gründen interessant sein. Oft sind insbesondere Nutzernetzwerke, gekoppelt an Text oder Multimediainhalte, von Interesse. Generell haben Onlineplattformen das Potential, Einblicke in Nutzeraktivitäten zu geben, etwa durch das Auslesen von Suchbegriffen, die von Nutzer/innen eingegeben werden, durch das Auswerten des Klickverhaltens auf verschiedene Links, durch das Aufdecken verschiedener Nutzernetzwerke oder durch das komplette Auswerten ganzer Textbeiträge, was beispielweise Einblicke in Meinungen und Stimmungen verspricht.

Manches daran ist neu, insbesondere die Vielfalt der Datenquellen, aber das grundlegende Prinzip erinnert stark an die Idee der prozessgenerierten Daten, die als nicht-reaktives Verfahren bereits ihren Platz in der sozialwissenschaftlichen Forschung gefunden haben, wie beispielsweise Daten zum Arbeitsmarkt, zur Einkommensstruktur, zur Mediennutzung, zum Bildungsstand. Dennoch wird im Kontext von Internetdaten auch oft von *New Data* gesprochen. Für Sozialwissenschaftler/innen werden Nutzungsdaten aus Internetportalen als eine mögliche neue Datenart angesehen, die – unabhängig von Einflüssen durch Studiendesigns – nicht nur Verhaltensweisen, sondern auch Meinungen offenlegen kann. Noch häufiger anzutreffen ist jedoch die Bezeichnung *Big Data*, die darauf anspielt, dass aus Internetdiensten große Menge von einzelnen Datenpunkten mit verhältnismäßig geringem Aufwand gewonnen werden können.

Von großen Datensätzen zu sprechen, ist zwar in vielen Fällen angebracht, dennoch driften die Meinungen darüber, ab wann eine Datenmenge als groß anzusehen ist, innerhalb der Forschungsgemeinschaft und vor allem auch zwischen den einzelnen Disziplinen auseinander: Für die einen ist alles groß, was den Rahmen der manuellen Inhaltsanalyse sprengt oder was nicht mehr in eine Excel-Tabelle passt, für andere fängt groß erst bei mehreren Terrabyte an und erfordert den Einsatz verteilter Rechnersysteme für die Speicherung und Auswertung der Daten. Kitchin und McArdle (2016) vergleichen 26 Big-Data-Datensätze und zeigen auf, wie schwierig es ist, allgemeingültige definitorische Kriterien für die Charakterisierung von Big Data festzulegen.

Die Frage, welche Art von Daten als Big Data bezeichnet werden, findet ganz unterschiedliche Auslegungen. Je nachdem, ob sie beispielsweise aus der Perspektive von Physiker/innen, Geograph/innen, Informatiker/innen, Geistes- oder Sozialwissenschaftler/innen betrachtet wird, umfasst die Bandbreite beispielsweise die Temperaturmessungen aller Wetterstationen über mehrere Jahre oder das gesamte Vokabular in Shakespeares Werken. Über Fächergrenzen hinweg gelten dabei Onlineumgebungen als interessante neue Datenquelle, die für ihre jeweiligen Fragestellungen neue Erkenntnisse versprechen (Kinder-Kurlanda/

https://doi.org/10.3224/84742233.12

Weller 2014: 96f). Aus sozialwissenschaftlicher Sicht ist das tatsächliche Datenvolumen mitunter eher nebensächlich. Entscheidender ist zunächst vielmehr die Frage nach der Datenqualität und der eigentlichen Aussagekraft von Datentypen, die ursprünglich nicht speziell für wissenschaftliche Fragestellungen gedacht waren und daher auf verschiedene Weise lückenhaft erscheinen können.

Da vielfach insbesondere die Nutzeraktivitäten und von Nutzer/innen generierte Inhalte wie Texte, Fotos und Videos als interessante Daten angesehen werden, sind sogenannte Social-Media-Plattformen eine Quelle für neuartige Forschungsdaten und deren Analyse. Hierzu zählen beispielsweise Dienste wie Facebook, Twitter, Instagram oder Foursquare sowie die Online-Enzyklopädie Wikipedia oder von Privatpersonen betriebene Blogs. Oft werden sie als eigener Forschungsgegenstand betrachtet. Jedoch spielen sie auch zunehmend in Kombination mit anderen Forschungsdaten eine Rolle, etwa für Vergleiche mit klassischen Medienanalysen oder als Ergänzung zu Umfragedaten. In diesem Kapitel geht es darum, Social-Media-Daten als eine Beispielmenge von Big bzw. New Data vorzustellen. Es sollen grundsätzliche Möglichkeiten der Forschung mit Social-Media-Daten aufgezeigt (Abschnitt 11.1 und 11.2), aber auch die bislang offenen Probleme der wissenschaftlichen Nutzung erläutert werden. Für Letzteres werden insbesondere die Datenqualität thematisiert (Abschnitt 11.3) sowie Probleme der Archivierung von Social-Media-basierten Forschungsdaten (Abschnitt 11.4) und drohender Datenverfall (11.5).

11.1 Social-Media-Daten als Forschungsgrundlage

Um die Frage zu beantworten, was alles als Social-Media-Daten zählt, müsste man zunächst definieren, was sich hinter Social Media verbirgt. Eine Definition von Social Media wird jedoch zunehmend schwieriger, denn eine Unterscheidung in *klassisches* Web und Social-Media-Plattformen als Verkörperung des *Mitmach-Webs* oder *Web 2.0* (O'Reilly 2005) ist heute kaum noch möglich.

Als diese Bezeichnungen vor mehr als zehn Jahren in Umlauf gebracht wurden, war es ohne besondere technische Kenntnisse kaum möglich, eigene Beiträge im Web zu veröffentlichen. Heute hingegen können innerhalb verschiedener Plattformen Bilder, Videos und Texte geteilt werden, es können Webinhalte kommentiert und bewertet und persönliche Beziehungen in Netzwerkplattformen abgebildet werden. Mitunter werden direkt aus dem Smartphone weitere Daten mit in die Onlineplattformen eingespeist – etwa der aktuelle Standort der Nutzer/innen, der in Form eines Geocodes einem Foto angehängt werden kann. Die Einbeziehung der Nutzer/innen in die Produktion von Webinhalten gilt als eines der Hauptmerkmale für Social-Media-Angebote.

Nach Schmidt (2009) spielt es darüber hinaus eine wesentliche Rolle, ob Nutzer/innen innerhalb einer Webplattform eigene Profile anlegen und pflegen. Diese dienen der Selbstdarstellung und der Vernetzung mit anderen Nutzer/innen innerhalb der Plattform. Anhand dieses Kriteriums ließen sich beispielsweise anonyme Kommentare in den Kommentarspalten von Onlinezeitungen oder Produktbewertungen in Einkaufsportalen ausklammern, da diese in der Regel nicht an ein individuelles Profil geknüpft sind und Nutzer/innen sich nicht untereinander vernetzen. Doch auch hier sind die Übergänge fließend. Die stetige Weiterentwicklung von Internetdiensten und Nutzungsgewohnheiten macht eine dauerhafte Definition von Social Media noch schwieriger.

Für die Nutzungsmöglichkeiten in der Forschung kommt es jedoch letztlich weniger auf eine präzise Definition an, als vielmehr darauf, zielsicher entscheiden zu können, welche

Datenquellen für die Beantwortung einer spezifischen Forschungsfrage in Betracht kommen. Hierzu muss man sich als Forscher/in stets einen guten Überblick über verschiedene Angebote und ihre Eigenschaften verschaffen, was dadurch erschwert wird, dass hier die Entwicklung oft rasant voranschreitet.

Interessant für die sozialwissenschaftliche Forschung können dabei grundsätzlich alle Plattformen sein, aus denen sich ein Erkenntnisgewinn über Verhalten oder Meinungen der Nutzer/innen ableiten lässt (Lazer/Radford 2017: 21f). Die Frage, ob dabei Aktivitäten an ein Nutzerprofil gebunden sind, wird vor allem dann interessant, wenn zumindest rudimentäre demographische Informationen oder die Position einer Person in einem Netzwerk mit ausgewertet werden sollen. Doch auch die Verfügbarkeit von Profilen bedeutet längst nicht, dass hieraus sinnvolle demographische Informationen abgeleitet werden können. Oft sind die Profilinformationen für sozialwissenschaftliche Ansprüche dürftig, Angaben zu Geschlecht, Wohnort, Alter oder Beruf sind selten akkurat verfügbar. Teilweise wird versucht, diese Informationen anderweitig abzuleiten, beispielsweise das Geschlecht auf Grundlage von Profilbildern oder Namensangaben (Karimi et al. 2016).

Zahlreiche Plattformen üben einen starken Reiz auf Wissenschaftler/innen verschiedener Disziplinen aus. Auf spezialisierten wissenschaftlichen Konferenzen im Bereich der Social-Media-Forschung, wie etwa die International Conference on Web and Social Media (ICWSM), die Social Media & Society Conference oder die Web Science Conference, findet man u.a. Studien zu bekannten Plattformen wie YouTube, Wikipedia, reddit, Tumblr, aber auch zu Dating-Plattformen, Online-Games oder Portalen zur Bewertung von Produkten. Besonders prominent in der Social-Media-Forschung sind jedoch die Plattformen Facebook und Twitter (Weller 2015: 285).

Untersucht werden beispielsweise politische Kommunikation (z.B. Jungherr/Schoen/Jürgens 2016), insbesondere auch im Kontext von Wahlprognosen (z.B. Metaxas/Mustafaraj/Gayo-Avello 2011), Gesundheit (z.B. Song/Gruzd 2017), Protest und Aktivismus (z.B. Jungherr/Jürgens 2014) oder Vertrauen und Diskriminierung (z.B. Edelman/Luca 2014). Lazer und Radford (2017) fassen derartige Forschungsansätze unter der Beschreibung *Digital Life* zusammen. Gemeint ist damit Forschung, die auf Onlineplattformen basiert, die zunehmend in alltägliche Lebensbereiche integriert sind und daher auch potentiell etwas über die Lebensweise der Nutzenden aussagen können.

Je nach Fragestellung stehen verschiedenste Datenformate im Vordergrund. Mal werden vor allem von Nutzenden verfasste Texte ausgewertet, mal rücken Fotos sowie andere Bilder oder auch Multimediadateien in den Fokus. In anderen Fällen sind es die Beziehungen zwischen den Nutzenden, die auf verschiedenen Interaktionen basieren können (z.B. explizite Verbindungen als Kontakte oder als *Freunde* bzw. *Follower* – aber auch implizitere Vernetzungen, z.B. basierend auf der Beteiligung an den gleichen Gesprächsthemen).

Davon abhängig, welche Daten untersucht werden sollen, kommen auch verschiedenste Methoden für die Datenanalyse in Betracht. Dazu zählen unterschiedlichste Formen der manuellen oder automatischen Textanalyse (darunter auch Sentiment Analysis oder Topic Modeling), sowie Netzwerkanalysen. Vielfach verfügen die Daten über zusätzliche Metadaten, insbesondere Zeitstempel und Geoinformationen, die weitere Datenanalysen ermöglichen. Einen Überblick über verschiedene methodische Herangehensweisen liefert z.B. das Handbuch *The Sage Handbook of Social Media Research Methods* von Sloan und Quan-Haase (2016).

Die Auswahl geeigneter Social-Media-Plattformen als Datenquellen für eine bestimmte Fragestellung ist der erste Schritt bei der Arbeit mit Social-Media-Daten. Sie steht damit am Anfang einer Reihe von grundlegenden Punkten, die interessierte Wissenschaftlerinnen und Wissenschaftler berücksichtigen sollten, bevor sie sich für die Arbeit mit Social-Media-Daten entscheiden. Schaukasten 11.1 fasst einige dieser Punkte in Form von Fragen zusammen.

Schaukasten 11.1: Fragen, die vor der Arbeit mit Social-Media-Daten beachtet werden sollten

Charakteristiken der Social-Media-Plattform

- Welche Social Media-Plattformen sind für mich relevant?
  - Warum ist eine bestimmte Plattform für meine Fragestellung relevant (z.B. Zielgruppe, Medienformat, Inhalte zu bestimmten Themen)?
  - Soll nur eine Plattform betrachtet oder sollen mehrere verglichen werden?
- Welche Dimensionen einer Social-Media-Plattform sind für mich interessant?
  - Sollen nur bestimmte Teilbereiche untersucht werden, beispielsweise nur Texte, Videos, Nutzernetzwerke?
- Sind Nutzungsstatistiken für die jeweilige Plattform verfügbar?
  - In welcher Form liegen Nutzerzahlen vor (z.B. Anteil der Gesamtbevölkerung oder Online-Bevölkerung, Unterscheidung in aktive und passive Nutzende, Angaben für tägliche, wöchentliche Nutzung)?
  - Ist die demographische Zusammensetzung der Nutzer/innen einer Plattform bekannt?

Datensammlung

- Wie können Daten abgerufen werden?
  - Sind Schnittstellen zum Datenabruf bei der Plattform selbst verfügbar (APIs)?
  - Können bereits bestehende Datensätze weiterverwendet werden?
  - Wird ein Datenzugriff über Drittanbieter angeboten? Wer bietet diesen an, wie vertrauenswürdig sind die Anbieter?
  - Ist der Datenzugriff kostenlos?
- Welche Daten sind erhältlich?
  - Nach welchen Kriterien können Daten ausgewählt werden (beispielsweise thematisch, nach Nutzergruppen, nach Regionen, nach Datum)? Sind die Suchkriterien für meine Forschungsfrage geeignet?
  - In welchem Dateiformat sollen die Daten vorliegen?
  - Ist der Zugriff beschränkt auf ein bestimmtes Datenvolumen? Welche Datenmenge wird benötigt?
  - Ist der Zugriff beschränkt auf bestimmte Zeiträume?
- Welcher Erhebungszeitraum soll gewählt werden?
  - Sollen über einen bestimmten Zeitraum alle Daten gesammelt oder Stichproben genommen werden?
  - Im Falle von Stichproben: Wie sollen diese angesetzt werden?
- Entstehen Verzerrungen durch die Datenauswahl?
  - Wird durch die Datenauswahl eine bestimmte Nutzergruppe bevorzugt (z.B. Vielnutzer, Nutzer/innen aus bestimmten Regionen)?
  - Kann der Zeitraum der Datensammlung eine Verzerrung hervorrufen (beispielsweise, wenn Feiertage oder Großereignisse die Nutzungsfrequenz beeinflussen)?

Forschungsethik

- Wird die Zustimmung der Nutzer/innen vorausgesetzt?
  - Willigen die Nutzer/innen über die Nutzungsbedingungen der Plattform dazu ein, dass ihre Daten an Dritte weitergegeben werden?
  - Werden nur öffentlich zugängliche Inhalte verwendet?
  - Kann man davon ausgehen, dass die Nutzer/innen sich der öffentlichen Einsehbarkeit ihrer Kommentare und Aktivitäten bewusst sind?
- Ist eine Anonymisierung möglich?
  - Können Nutzer/innen anhand vorhandener Informationen wieder identifiziert werden (etwa durch eine Textsuche bei vorliegenden Zitaten aus Tweets)?
  - Handelt es sich um Nutzergruppen, die besonders auf Anonymität angewiesen sind (z.B. politische Aktivist/innen, Minderjährige) oder um Nutzergruppen, die eine namentliche Erwähnung sogar bevorzugen würden (z.B. Autor/innen, Künstler/innen)?

Nachnutzbarkeit

- Sollen die Daten nach der Nutzung Dritten zugänglich gemacht werden?

- Gibt es rechtliche oder technische Rahmenbedingungen, die Einfluss darauf haben, welche Daten weitergegeben werden müssen?
- Werden beim Datenkauf Nutzerbedingungen unterschrieben, die eine Weitergabe an Dritte einschränken oder gar untersagen?

Quelle: Eigene Darstellung

Die folgenden Abschnitte dieses Kapitel befassen sich mit diesen Aspekten. Dabei steht der Bereich der Datensammlung (Abschnitt 11.2) im Vordergrund. Generell ist zu berücksichtigen, dass die im Schaukasten angesprochenen Faktoren die Qualität der Daten bzw. ihre Brauchbarkeit für eine bestimmte Forschungsfrage beeinflussen können.

11.2 Möglichkeiten und Grenzen der Datensammlung im Social Web

Je nach Plattform sind Social-Media-Daten unterschiedlich gut für die Sammlung und Nutzung für die Forschung zugänglich. Da es sich bei den Plattformen in der Regel um kommerzielle Dienste handelt, deren Betreiber eigene wirtschaftliche Interessen verfolgen, ist ein offener Datenzugang, etwa für wissenschaftliche Zwecke, selten vorgesehen. Eine Ausnahme stellt Wikipedia dar. Routinemäßig werden Kopien, sogenannte Wikipedia Dumps, der kompletten aktuellen Version der Community-basierten Online-Enzyklopädie für die Nachnutzung zur Verfügung gestellt.

In einigen anderen Fällen verfügen Social-Media-Plattformen über spezielle Schnittstellen, über die in gewissem Umfang Daten abgerufen werden können. Eine solche Schnittstelle, genannt *Application Programming Interface* (API), dient in erster Linie jedoch dazu, die jeweilige Plattform für die Verbindung mit anderen Anbietern nutzbar zu machen. So kann etwa damit ein auf Instagram hochgeladenes Foto direkt bei Facebook geteilt werden. Der Funktionsumfang und die Nutzungsbedingungen sind für eine derartige Nutzung ausgerichtet. Dass auch Wissenschaftler/innen die API verwenden, um Daten zu sammeln, ist aus Anbietersicht wohl eher ein Nebeneffekt. Man sollte sich deshalb stets dessen bewusst machen, dass die auf diesem Wege abrufbaren Daten in keiner Weise speziell für die wissenschaftliche Datenerhebung aufbereitet wurden. Die Qualität der Daten ist, wie wir unten sehen werden, mitunter unsicher. Zudem ist der Zugang oft nur mit entsprechenden technischen Vorkenntnissen möglich und das Zugriffsvolumen in der Regel begrenzt. Oft sind die Daten in bestimmten Formaten abrufbar, was einerseits eine gute Strukturierung mit sich bringt, andererseits aber auch dazu führen kann, dass die Daten *unvollständig* sind. So fehlen beispielsweise bei Twitter-Daten im textbasierten JSON-Format die in den ursprünglichen Tweets enthaltenen Videos oder Bilder (vgl. Abbildung 11.1).

Drittanbieter haben früh einen Markt darin gesehen, Tools für das Auslesen von Social-Media-APIs zur Verfügung zu stellen, mit denen Interessierte ohne eigene Programmierkenntnisse die Daten auslesen konnten. Insbesondere für Twitter existierten verschiedene Angebote, bei denen innerhalb einer Weboberfläche mit wenigen Klicks eine eigene Datensammlung aufgesetzt und die Daten später z.B. als Excel-Datei heruntergeladen werden konnten. In bestimmten Wissenschaftskreisen war beispielsweise die Plattform TwapperKeeper besonders beliebt. Sie ermöglichte es, Tweets zu bestimmten Schlagworten oder Hashtags zu sammeln und die Sammlungen wiederum anderen Nutzer/innen zugänglich zu machen. Im Jahr 2011 gab es jedoch bei Twitter eine größere Umstellung der APIs und deren Nutzungsbedingungen, in deren Folge viele dieser Dienste nicht weiterbetrieben werden durften (Bruns 2011). Auch TwapperKeeper musste seinen Service schließen, bot jedoch wenig später mit YourTwapperKeeper eine überarbeitete Version an. Diese steht allerdings

nicht länger als Webinterface zur Verfügung, sondern muss von den Nutzer/innen auf ihrem eigenen Server installiert werden (Bruns/Liang 2012). Dieses Beispiel verdeutlicht, dass die verfügbaren Datenzugänge Änderungen unterworfen sind, die von den Plattformbetreibern kurzfristig eingeführt werden können. Neben praktischen Konsequenzen für die Datensammlung können Änderungen beispielsweise der API-Nutzungsbedingungen auch Auswirkungen auf die Datenqualität oder die Vergleichbarkeit der Daten über größere Zeiträume hinweg haben.

Für den Zugang zu Twitter-Daten können Interessierte aktuell aus verschiedenen unterstützenden Tools[1] wählen, die teilweise von Universitäten, wie z.B. COSMOS (Burnap et al. 2015) oder Social Feed Manager, teilweise kostenpflichtig von Unternehmen angeboten werden, wie etwa Tweet Archivist oder DiscoverText. Zudem wurden Plugins für etablierte Analysesoftware entwickelt, wie etwa NVIVO, die einen direkten Abruf von Twitter-Daten ermöglichen. Einen Überblick über die verschiedenen Optionen gibt beispielsweise Littman (2017a). Da jedoch alle diese Dienste letztlich auf den Twitter-APIs basieren, sind sie insgesamt auch den Einschränkungen unterworfen, die Twitter für die APIs allgemein auferlegt hat (Gaffney/Puschmann 2014). Twitter bietet den Datenabruf über verschiedene APIs an, die aber jeweils nur einen Ausschnitt des gesamten Twitter-Volumens zugänglich machen. Darüber hinaus ist es beispielsweise nicht möglich, rückwirkend alle Tweets zu einem bestimmten Suchbegriff oder Hashtag abzurufen. Dementsprechend muss man in der Regel im Voraus planen, welche Art Daten man abrufen möchte. Das funktioniert relativ gut, wenn man Informationen über ein vorab geplantes Ereignis sammeln möchte, etwa für eine Wahl.

Bei spontanen Ereignissen, wie etwa Protestbewegungen oder Naturkatastrophen, stellt sich die voraus geplante Datensammlung jedoch weit schwieriger dar. Wenn man heute beispielsweise rückwirkend alle Tweets der Occupy-Wallstreet-Proteste basierend auf den zugehörigen Hashtags abrufen will, helfen die frei zugänglichen Twitter-APIs nicht weiter. Den Vollzugriff auf die gesamte Datenbasis und auf sogenannten *historischen* Tweets (darunter versteht Twitter alle Tweets, die rückwirkend erfasst werden sollen) vermarktet Twitter kostenpflichtig. Das Preismodell liegt auf einem Level, das den Einkauf für Wissenschaftler/innen nicht immer spontan erschwinglich macht. Die Preise sind von verschiedenen Faktoren in Zusammenhang mit der Suchanfrage und dem Tweet-Volumen abhängig, ein aktueller Beispielwert wäre etwas unter 2.000 Dollar für rund 1,5 Millionen Tweets. Für den Vertragsabschluss brauchen Forscher/innen zudem in der Regel die Unterschrift der Instituts-/Universitätsleitung. Zunehmend findet man Beispiele von Forscher/innen, die sich einen solchen Datenankauf für ihre Forschungsarbeiten ermöglicht haben. Eine Anlaufstelle für den Einkauf verschiedener Social-Media-Inhalte ist beispielsweise DataSift..

Sowohl die APIs und darauf basierende Tools als auch die kostenpflichtigen Angebote liefern in erster Linie textbasierte Daten in standardisierten Formaten, wie etwa das textbasierte JSON-Format für Twitter-Daten. Abbildung 11.1 zeigt als Beispiel den ersten Tweet des Twitter-Mitgründers Jack Dorsey im JSON-Format. Das Format besteht aus fixen Metadatenelementen und enthält so beispielsweise Angaben zum Veröffentlichungszeitpunkt (*created_at*: *Tue Mar 21 20:50:14 +0000 2006*), zur Sprache (*lang*: *En*) oder zum Autor[2] (*name*: *Jack Dorsey*) bzw. zum Nutzernamen bei Twitter (*screen_name*: *Jack*). Alle Angaben sind nach einem vorgegebenen Muster strukturiert verfügbar, jedoch rein textbasiert. Möchte man über den Standard hinausgehen und beispielsweise bei Twitter eingebettete Fotos oder

1 Manche der im Folgenden genannten Tools ermöglichen auch den Zugriff auf andere Social-Media-Daten und sind nicht auf Twitter beschränkt.

2 Die Angaben zur Autorin/zum Autor eines Tweets beziehen sich jeweils auf die verfügbare Angabe im Nutzerprofil. Nutzerinnen und Nutzer müssen hier jedoch keinen echten oder vollständigen Namen angeben.

Videos ebenfalls auswerten, so muss man in der Regel eigene Tools bauen, die speziell auf diese Arten von Inhalten ausgerichtet sind.

Abbildung 11.1: Beispiel für einen Tweet im JSON-Format


```
{
  "created_at": "Tue Mar 21 20:50:14 +0000 2006",
  "id": 20,
  "id_str": "20",
  "text": "just setting up my twttr",
  "source": "web",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 12,
    "id_str": "12",
    "name": "Jack Dorsey",
    "screen_name": "jack",
    "location": "California",
    "description": "",
    "url": null,
    "entities": {
      "description": {
        "urls": []
      }
    },
    "protected": false,
    "followers_count": 2577282,
    "friends_count": 1085,
    "listed_count": 23163,
    "created_at": "Tue Mar 21 20:50:14 +0000 2006",
    "favourites_count": 2449,
    "utc_offset": -25200,
    "time_zone": "Pacific Time (US & Canada)",
    "geo_enabled": true,
    "verified": true,
    "statuses_count": 14447,
    "lang": "en",
    "contributors_enabled": false,
    "is_translator": false,
    "is_translation_enabled": false,
    "profile_background_color": "EBEBEB",
    "profile_background_image_url": "http://abs.twimg.com/images/themes/theme7/bg.gif",
    "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme7/bg.gif",
    "profile_background_tile": false,
    "profile_image_url": "http://pbs.twimg.com/profile_images/448483168580947968/pL4ejHy4_normal.jpeg",
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/448483168580947968/pL4ejHy4_normal.jpeg",
    "profile_banner_url": "https://pbs.twimg.com/profile_banners/12/1347981542",
    "profile_link_color": "990000",
    "profile_sidebar_border_color": "DFDFDF",
    "profile_sidebar_fill_color": "F3F3F3",
    "profile_text_color": "333333",
    "profile_use_background_image": true,
    "default_profile": false,
    "default_profile_image": false,
    "following": true,
    "follow_request_sent": false,
    "notifications": false
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "retweet_count": 23936,
  "favorite_count": 21879,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "urls": [],
    "user_mentions": []
  },
  "favorited": false,
  "retweeted": false,
  "lang": "en"
}
```


Quelle: Der erste Tweet des Twitter-Mitgründers Jack Dorsey (@jack) aus dem Jahr 2006. Der Originaltweet ist verfügbar unter: https://twitter.com/jack/status/20.

Neben den kommerziellen Interessen der Anbieter spielen schließlich auch die Privatsphäre-Einstellungen der Social-Media-Nutzer/in eine entscheidende Rolle dabei, welche Arten von Daten für die (wissenschaftliche) Nutzung frei abrufbar sind. So sind Twitter-Daten sicher auch deswegen verhältnismäßig beliebt als Forschungsgrundlage, weil der überwiegende Großteil der Twitter-Inhalte öffentlich zugänglich ist. Als Twitter-Nutzer/in hat man

lediglich die Wahl zwischen zwei Privatsphäre-Einstellungen: Das komplette Profil und alle Tweets sind entweder komplett öffentlich – und über die oben genannten Datenzugänge kommt man bei Twitter immer nur an den öffentlichen Teil der Daten – oder komplett privat und damit nur für einzeln zugelassene Nutzer/innen einsehbar (Zimmer/Proferes 2014a).

Bei Facebook beispielsweise ist die Lage deutlich komplexer. Ursprünglich regelte die dortige Standardeinstellung, dass Beiträge nur für Freunde – ggf. auch noch für deren Freunde – sichtbar waren. Inzwischen können Nutzer/innen für jeden einzelnen Beitrag eine andere Sichtbarkeit einstellen, z.B. nur für bestimmte Freunde, alle Freunde oder komplett öffentlich. Dadurch sind einerseits anteilig weniger Daten aus Facebook öffentlich einsehbar als Daten aus Twitter. Andererseits stellt sich bei einigen öffentlichen Kommentaren auf Facebook die Frage, ob sich Nutzende in diesem Fall der Öffentlichkeit ihrer Aussagen überhaupt bewusst waren oder ob sie sich selbst noch im geschützten Bereich wähnten. Ein Beispiel für einen solchen Fall von öffentlich einsehbaren Nutzungsaktivitäten findet man im Kommentarbereich von Angela Merkels Facebook-Seite. Die Kommentare sind ohne Facebook-Login für jeden öffentlich einsehbar, doch es ist unklar, ob das den Kommentierenden bewusst ist.

Der Umgang mit solchen und ähnlichen Social-Media-Inhalten fällt bislang noch in einen großen Graubereich: Sowohl der rechtliche Rahmen als auch Fragen der Forschungsethik sind meist nicht vollständig geklärt. Interviews mit Social-Media-Forscher/innen zeigen, dass zwar ein Bewusstsein für die Bedeutung von Forschungsethik in diesem Bereich vorhanden ist, dass aber in der Forschungsgemeinschaft bisher keine Einigkeit darüber besteht, wie genau der ethisch verantwortungsbewusste Umgang mit verschiedenen Datentypen aussieht (Weller/Kinder-Kurlanda 2014: 1). Selbst bei den öffentlich vorliegenden Twitter-Daten gehen hierzu die Meinungen auseinander. Und die bisher verfügbaren Richtlinien verschiedener Fachgemeinschafen bieten vorwiegend allgemeine Denkanstöße, aber keine detaillierten Anweisungen für konkrete Einzelfälle. Als Beispiel seien hier vor allem die Guidelines der Association of Internet Researchers (AoIR) genannt (Markham/Buchanan 2012). Einen relativ breiten Überblick über aktuelle Herausforderungen im Bereich Forschungsethik bei Internetdaten und Beispielszenarien liefern Zimmer und Kinder-Kurlanda (2017).

In Bezug auf den ethisch-bewussten Umgang mit Forschungsdaten aus Social-Media-Umgebungen ist langfristig die Etablierung von Standards im Sinne guter wissenschaftlicher Praxis notwendig. Hierbei können und sollten Datenarchive künftig eine führende Rolle übernehmen, da sie über langjährige Erfahrungen beim Management unterschiedlichster Datentypen und der Entwicklung entsprechender Standards verfügen. Archivierte Datensätze, die bereits auf Einhaltung ethischer Standards geprüft wurden, könnten dann wiederum für andere Forschende zugänglich gemacht werden, die somit nicht ihrerseits von vorne mit diesbezüglichen Überlegungen beginnen müssten.

Allgemein wären Fortschritte im Bereich der Datenarchivierung von Social-Media-Daten äußerst wünschenswert (Weller/Kinder-Kurlanda 2016). Sie könnten dabei helfen, den Datenzugang insgesamt zu erleichtern und damit ggf. auch aktuelle Ungleichgewichtungen relativieren zwischen den Wissenschaftler/innen, die sich kostenpflichtige Vollzugriffe auf Daten leisten können oder durch persönliche Kontakte zu Social-Media-Firmen über privilegierte Zugriffsmöglichkeiten verfügen und der breiteren Masse der an Social-Media-Daten für ihre Forschung Interessierten. Solche Ungleichheiten werden seit einiger Zeit in der Social-Media- und Big-Data-Forschung kritisiert (Boyd/Crawford 2012: 673ff).

Der Zugang zu Social-Media-Daten erfordert oftmals einen hohen Aufwand für die Einarbeitung in die Umsetzung der Datensammlung, zumal sich die Rahmenbedingungen für die Zugänglichkeit bei einzelnen Plattformen rasch ändern können. Für besonders relevante Themen ist davon auszugehen, dass verschiedene Forschergruppen parallel Zeit und Aufwand in die Erstellung relativ ähnlicher Datensätze stecken. So gibt es beispielsweise mindestens 17

verschiedene Datensätze basierend auf Twitter-Daten rund um die Präsidentschaftswahl in den USA in 2012 (Weller 2014: 246). Hier könnten zentral archivierte Datensätze in Zukunft idealerweise einen vermeidbaren Mehraufwand reduzieren. Unter Umständen können Archivierungslösungen zudem langfristig dazu beitragen, die Qualität von Social-Media-Datensätzen zu verbessern, insbesondere bezogen auf Transparenz und Reproduzierbarkeit der damit durchgeführten Studien.

11.3 Datenqualität von Social Media Daten

Social-Media-Daten üben also für mehr und mehr Forscher/innen als neue Datentypen einen gewissen Reiz aus: Sie sind schnell verfügbar, bilden potentiell das Verhalten von Nutzenden auf der ganzen Welt ab, und liegen meist in strukturierter Form vor. Wie wir gesehen haben, ist jedoch bereits der Datenzugriff beschränkt und nicht auf eine wissenschaftliche Nutzung ausgerichtet. Auch die Datenqualität entspricht nicht unbedingt wissenschaftlichen Erwartungen. Eine Herausforderung für die Datenqualität ist es, ihre Repräsentativität sinnvoll einzuschätzen bzw. etwaige Verzerrungen (*Biases*) zu erkennen. Denn wie auch Lazer et al. (2014: 1203) warnen: „The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis."

Der Datenzugriff und die bereits erwähnten etwaigen Beschränkungen der Plattformbetreiber sind eine Stelle, an der es zu Verzerrungen kommen kann. Morstatter et al. (2013) sowie Morstatter, Pfeffer und Liu (2014) untersuchen beispielsweise, wie sich der über die APIs frei verfügbare Auszug aus Twitter zu dem kostenpflichtigen Vollzugriff verhält und weisen auf diese Weise u.a. Verzerrungen in Bezug auf die thematische Abdeckung nach.

Weitere Verzerrungen können in Bezug auf die Bevölkerungsrepräsentation entstehen. In den seltensten Fällen ist davon auszugehen, dass Social-Media-Daten für bestimmte Bevölkerungsgruppen repräsentativ sind. Umso wichtiger ist es, einzuschätzen, wie sich die Nutzerschaft einer bestimmten Social-Media-Plattform zusammensetzt. Im ersten Schritt ist herauszufinden, welcher Anteil der Bevölkerung eine Plattform in welchem Umfang nutzt.[3] Auf dieser Basis kann die gesellschaftliche Rolle einer Plattform eingeordnet werden. So ist Facebook zwar in einigen Ländern höchst populär, wird in anderen dafür kaum genutzt – und ist für Studien zur dortigen Bevölkerung somit als Forschungsdatenquelle weit weniger geeignet. Nichtnutzung bestimmter Plattformen kann z.B. an Zugriffsbeschränkungen und Zensuren liegen, etwa in China. Es kann aber auch sein, dass andere Plattformen, wie der Facebook-Konkurrent Vkontakte in Russland, populärer sind. Nicht immer ist es möglich, genauere Informationen zur Nutzung nach demographischen Merkmalen zu erhalten. Am ehesten sind Angaben zu dem Alter des Anteils der Bevölkerung, die einen Dienst nutzt, erfassbar. Wünschenswert wären mitunter auch Angaben zu Geschlecht, Einkommen oder Bildungsstand. Blank (2017: 680ff) gibt einen guten Überblick über Versuche, die Zusammensetzung der Gruppe der Twitter-Nutzer/innen automatisch zu entschlüsseln. So sollen z.B. anhand von Vornamen, Ortsangaben oder Fotos das Geschlecht, Alter oder Herkunft von Twitter-Nutzenden ermittelt werden. Blank (ebd.: 681) weist zudem darauf hin, dass es nicht reicht, die Zusammensetzung der Nutzerschaft einer Plattform zu kennen. Relevant ist im Weiteren die Kenntnis darüber, inwiefern diese sich von den Nutzenden anderer Plattformen, der

3 In Deutschland nutzen in 2017 nach Angaben von Koch und Frees (2017) beispielsweise 3 % der Bevölkerung Twitter und 33 % Facebook mindestens einmal wöchentlich.

Online-Bevölkerung oder der Offline-Bevölkerung unterscheidet. Mit Hilfe von Umfragen vergleicht er US-amerikanische und britische Twitter-Nutzer/innen mit anderen Gruppen, etwa den Nicht-Twitter-Nutzer/innen und der Offline-Bevölkerung. Er zeigt u.a., dass britische Twitter-Nutzer/innen jünger, wohlhabender und besser ausgebildet sind als andere Internetnutzer (ebd.: 683ff). Solche Unterschiede sind insbesondere relevant, wenn man Social-Media-Daten für Prognosen (z.B. von Wahlergebnissen oder Kinoerfolgen) nutzen möchte. Blank und Lutz (2017) betrachten verschiedene Social-Media-Plattformen in Bezug auf Repräsentativität für Nutzergruppen. Darüber hinaus werden aktuell erste Versuche mit speziell in Panels rekrutierten Nutzer/innen für die Erfassung der Social-Media-Nutzung erprobt (Resnick/Adar/Lampe 2015).

Bruns und Stieglitz (2014: 241ff) weisen zudem darauf hin, dass die Repräsentativität von Social-Media-Daten an mehreren Stellen kritisch werden kann. Wie gut die Gesamtbevölkerung unter den Plattformnutzenden repräsentiert ist, ist nur der Anfang. Man muss auch fragen, inwieweit ein bestimmter aus einer Plattform extrahierter Datensatz für diese Plattform repräsentativ ist. Twitter-Daten werden beispielsweise häufig basierend auf Hashtags zusammengestellt (z.B. alle Tweets mit dem Hashtag *#btw17* für einen Datensatz zur Bundestagswahl 2017). Hashtags werden allerdings nicht von allen Twitter-Nutzer/innen gleichermaßen verwendet. Bei der Datensammlung basierend auf Hashtags ist davon auszugehen, dass insbesondere die Aktivitäten von erfahrenen Twitter-Nutzer/innen gemessen werden, weniger die von Twitter-Neulingen oder gelegentlich Nutzenden (ebd.: 241). Eine weitere Herausforderung ist die Verwendung verschiedener Hashtags für das gleiche Thema, bei der Bundestagswahl 2017 z.B. *#btw17* oder *#bundestagswahl*. Bei Hashtags handelt es sich nicht um kontrolliertes Vokabular, sie können frei gewählt werden. In manchen Fällen verwenden bestimmte Nutzergruppen gezielt eher das eine, andere Nutzergruppen ein anderes Hashtag, teils versehentlich, etwa wegen verschiedener Muttersprachen, teils um unter sich zu bleiben und sich abzugrenzen. Beschränkt man in solchen Fällen die Datensammlung auf eines der beiden Hashtags, beeinflusst dies die betrachtete Nutzergruppe. Auch Geoinformationen, die an Tweets angehängt werden können, sind als Kriterium für die Datensammlung problematisch, da nur ein sehr geringer Teil der Twitter-Nutzenden diese Geocodes verwendet (Sloan/Morgan 2015: 2). Ruths und Pfeffer (2014) sehen ein weiteres Repräsentativitätsproblem darin, dass die Aktionen innerhalb von Social-Media-Plattformen nicht unbedingt mit den scheinbar repräsentierten Offline-Aktivitäten übereinstimmen. So ist ein *Freund* auf Facebook nicht unbedingt ein Freund im Offline-Leben. Für Forscherende liegt also die Herausforderung auch darin, herauszufinden, welche Aktionen innerhalb einer Plattform dem entsprechen könnten, was sie mit Hilfe der Social-Media-Daten messen wollen. Als eine weitere Schwierigkeit kommt für die Forschung hinzu, dass sich sowohl die Social-Media-Plattformen (Karpf 2012: 643f) als auch die Nutzungspraktiken rasant weiterentwickeln.

Aussagen zur Qualität von Social-Media-Daten werden zudem dadurch erschwert, dass die in aktuellen Forschungsarbeiten verwendeten Daten nur in Ausnahmefällen für die Nachnutzung verfügbar gemacht werden. So ist es oft nicht möglich, Forschungsergebnisse zu replizieren oder zu überprüfen, wie in Kapitel 7.1 diskutiert. Dieser Zustand ist möglicherweise ein Ausdruck dessen, dass sich zur Erforschung von Social Media bisher keine spezialisierte Fachdisziplin etabliert hat und vielmehr in verschiedensten Disziplinen mit Ansätzen und Methoden experimentiert wird. So steht derzeit noch die Exploration von Daten zu einem bestimmten Thema im Vordergrund. Langfristig wird die Social-Media-Forschung hier jedoch eigene Kriterien zur Qualitätsbewertung von Datensammlung und Datendokumentation und einheitliche Standards entwickeln müssen, um den wissenschaftlichen Erkenntnisgewinn voranzubringen. Ein Wandel in diese Richtung deutet sich bereits an; es werden zunehmend Qualitätsprobleme kritisiert (z.B. Lazer/Radford 2017; Resnick/Adar/Lampe 2015; Schroeder 2014; Tufekci 2014).

Fehlende Reproduzierbarkeit der in der Social-Media-Forschung erzielten Ergebnisse kann sich langfristig auf das gesamte Forschungsfeld sehr negativ auswirken. Erste Fachzeitschriften und Konferenzveranstalter suchen deswegen nach Möglichkeiten, die zugrunde liegenden Datensätze gemeinsam mit den angenommenen wissenschaftlichen Veröffentlichungen bereitzustellen. Auf der International Conference on Web and Social Media (ICWSM) wird den Autor/innen der angenommenen Beiträge beispielsweise angeboten, Ihre Datensätze gleich mit zu publizieren (ICWSM 2012).

Hinzu kommt bei dieser Konferenzreihe seit Kurzem auch eine neue Kategorie an Beiträgen, sogenannte Dataset Papers, die sich allein der Publikation eines Datensatzes mit dazugehörender detaillierter Beschreibung desselben widmen (vgl. 7.3.2). Der Bedarf, Daten zu teilen und zu archivieren, ist also vorhanden, was auch Interviews mit Social-Media-Forscher/innen bestätigen (Weller/Kinder-Kurlanda 2015: 31f). Schwierig wird es jedoch im Detail in der Umsetzung, da sich traditionelle Modelle der Datenarchivierung nicht immer eins zu eins übertragen lassen.

11.4 Archivierung und Nachnutzung von Social-Media-Daten

Bisher gehen Ansätze zur Archivierung von Social-Media-Daten von ganz unterschiedlichen Gruppen aus. Verlage und Konferenzveranstalter sind, wie wir eben gesehen haben, eine solche Interessensgruppe. Daneben finden wir auch verschiedene Eigeninitiativen einzelner Wissenschaftler/innen oder Arbeitsgruppen. Vieles davon ist improvisiert und geschieht mit wenig Bezug zu etablierten Praktiken aus dem Bereich Forschungsdatenmanagement. Initiativen von auf Archivierung spezialisierten Expert/innen und Einrichtungen sind bislang eher die Ausnahme. Einen Überblick über bestehende Ansätze liefern Thomson (2016) sowie Weller und Kinder-Kurlanda (2016). Einzelne Beispiele sollen im Folgenden kurz skizziert werden:

- Neben der ICWSM veröffentlichen auch andere Konferenzreihen Datensätze – mitunter geknüpft an sogenannte Data Challenges. Dabei wird ein Datensatz vor der Konferenz zur Verfügung gestellt und mit einer Aufgabe verbunden, wie z.B. bestimmte Informationen daraus zu extrahieren. Die Wissenschaftler/innen, die sich dieser Aufgabe stellen, können sich anschließend bei der Konferenz über ihre Ansätze austauschen. Ein Beispiel hierfür ist die Text Retrieval Conference (TREC), die mehrfach Twitter-Datensätze für Aufgaben zum Information Retrieval[4] bereitgestellt hat (TREC 2011).
- Einzelne Wissenschaftler/innen teilen Datensätze z.T. über ihre eigenen oder institutionellen Websites. Dies hat jedoch auch schon zu Fällen geführt, in denen der Datensatz im Laufe der Zeit nicht mehr über die Website verfügbar war. So geschehen im Falle von Cha et al. (2010), wo der zum Paper gehörende Datensatz auf Anweisung von Twitter wieder von der Webseite des Max Planck Institute for Software Systems entfernt werden musste.
- Andere Wissenschaftler/innen haben ihre Datensätze auch bereits über Archivierungsinstitutionen bereitgestellt, beispielsweise Summers (2014) über das Internet Archive oder Kaczmirek und Mayr (2015) über das GESIS Datenarchiv. In beiden Fällen handelt es sich um Twitter-Daten. Unter Berücksichtigung der Twitter-Nutzungsbedingungen werden dabei keine Texte oder JSON-Dateien, sondern Listen von Tweet-Identifikationsnummern (Tweet-IDs) archiviert und geteilt. Basierend auf den Tweet-IDs kann dann jeweils der Originaltweet wieder aufgerufen werden, sofern dieser nicht zwischenzeitlich gelöscht wurde. Bei GESIS liegen weitere Twitter-Datensätze in Form von Listen

4 Im Rahmen der Konferenz werden Aufgaben gestellt, mit Information Retrieval Algorithmen bestimmte Informationen aus Twitter-Texten zu identifizieren, z.B. alle Eigennamen finden oder Ereignisse identifizieren. Die Konferenzveranstalter stellen einen Twitter-Datensatz als Bezugsrahmen zur Verfügung. Forscher/innen wenden ihre Retrieval-Verfahren auf diesen Datensatz an und reichen die damit erzielten Ergebnisse zum Vergleich bei den Veranstaltern ein.

von Tweet-IDs vor, die mit zusätzlichen Informationen wie Geocodierungen (Kinder-Kurlanda et al. 2017; Pfeffer/Morstatter 2016) oder Kategorisierungen (Nishioka/Scherp/Dellschaft 2015) angereichert wurden.

- Teilweise bauen Forschungsgruppen oder Forschungsinstitute ganze Sammlungen von Datensätzen auf. Beispiele hierfür sind das Projekt KONECT an der Universität Koblenz-Landau, das an Datensätzen speziell vom Typ Netzwerkdaten interessiert ist, oder das CrisisLex Projekt (Olteanu et al. 2014), das thematisch auf Krisenkommunikation bezogene Datensätze sammelt.
- Eine ganz besondere Situation gibt es zudem im Fall Twitter. Das Unternehmen Twitter Inc. hat selbst eine Initiative zur Archivierung seiner Inhalte in die Wege geleitet und bereits 2010 ein diesbezügliches Abkommen mit der Library of Congress in den USA getroffen (Stone 2010). Bis heute hat dieses Abkommen jedoch nicht zu einer praktischen Lösung geführt – ein Zugriff auf die bei der Library of Congress archivierten Twitter-Daten ist bislang nicht möglich und wenig Offizielles ist über den Stand der Entwicklungen bekannt (McLemee 2015; Zimmer 2015). Ende 2017 wurde zudem verkündet, dass statt der ursprünglich geplanten Vollarchivierung aller Tweets von nun an nur noch zu besonderen Themen Datensammlungen archiviert werden sollen (Osterberg 2017).

Darüber hinaus ist davon auszugehen, dass Datensätze oftmals sehr informell zwischen einzelnen Wissenschaftler/innen weitergegeben werden – über persönliche Kontakte und auf Anfrage in einer Art *grey market* (Weller/Kinder-Kurlanda 2015: 33). Mitunter gibt es auch Fälle, in denen große Datensätze zu Social-Media-Plattformen von Einzelpersonen im Internet zur Verfügung gestellt werden, ohne dass die Initiatoren und rechtlichen Status der Daten konkret bekannt sind. So gibt es beispielsweise einen reddit-Datensatz, der von reddit-Nutzer Jason Baumgartner veröffentlicht wurde (stuck_in_the_matrix 2015a und 2015b). Der Datensatz wurde kürzlich von Wissenschaftlern als unvollständig kritisiert (Gaffney/Matias 2018) und wird seitdem von verschiedenen Beteiligten inklusive Jason Baumgartner ergänzt und verbessert.

Bislang liegen wenige Informationen vor, in welchem Umfang die bereits zur Verfügung gestellten Datensätze auch von Dritten nachgenutzt werden. Als Zimmer und Proferes (2014b) eine Sammlung von Studien, die auf Twitter-Daten basieren, bezüglich ihrer Methoden und Herangehensweisen klassifizierten, verzeichneten sie zwischen 2010 und 2012 in 3–8 % der Fälle eine Nachnutzung bereits vorhandener Datensätze. In Interviews gaben manche Social-Media-Forscher/innen an, dass sie die von anderen zusammengestellten Daten lieber nicht nutzen würden, da sie dabei nicht nachvollziehen können, wie die Datensätze zustande gekommen sind (Weller/Kinder-Kurlanda 2015: 33f). Wenngleich sich diese Aussage sicher vorwiegend auf informell zwischen Kolleg/innen weitergereichte Datensätze bezieht, so ist es doch ganz allgemein um die Dokumentation der archivierten Daten noch spärlich bestellt. In den oben skizzierten Fällen wird auch im Bereich Datendokumentation vielfach improvisiert. Mitunter ist weder klar, wie die Archivierung eines bestimmten Datensatzes abgelaufen ist, noch bekannt, wie der Datensatz selbst generiert wurde. Und auch in Fällen, bei denen professionelle Archive beteiligt sind, existieren oft zunächst Behelfslösungen: Dokumentationsstandards, wie in Kapitel 9 beschrieben, sind auf Social-Media-Daten als neuen Datentyp noch nicht eingestellt. Ähnliches gilt für andere neue Datentypen, wie die in Kapitel 12.4 beleuchteten georeferenzierten Daten. Auch im Bereich der Forschungsethik gibt es noch viele offene Fragen. Langfristig ist zu hoffen, dass Archive eine erweiterte Expertise in Bezug auf Datennutzungsrechte und Datenschutz aufbauen können, von der alle profitieren würden. Das Inter-University Consortium for Political and Social Research (ICPSR) kündigte diesbezüglich jüngst den Aufbau eines eigenen Social-Media-Datenarchivs an (Hemphill/Leonard/Hedstrom 2018).

Problematisch ist für eine professionelle Archivierung auch, dass die Verschiedenheit von Social-Media-Daten einheitliche Regelungen erschwert. Aus Social-Media-Diensten erhobene Daten können je nach Forschungsfrage und Datensammlungsansatz für qualitative oder quantitative Forschung genutzt werden. Es können Texte, Bilder oder Multimediadateien,

Netzwerkdaten oder Mischungen aus all diesen vorliegen. Daher ist auch nicht unbedingt vorgegeben, welche Art Datenarchiv sich für diese Daten zuständig fühlen könnte und welche Best Practices auf den Umgang mit Social-Media-Daten übernommen werden sollten.

Dennoch lassen sich erste Ratschläge geben, worauf Forschende im Umgang mit Social-Media-Daten für eine potentielle Dokumentation und Archivierung achten sollten. Beispielsweise wird zunehmend zusätzlich zu den eigentlichen Daten der Programmiercode, mit dem die Datensammlung und -bereinigung durchgeführt wurde, gesichert oder geteilt. Dies trägt zur besseren Nachvollziehbarkeit der Datenzusammensetzung bei. Zu weiteren entsprechenden Maßnahmen zählen genaue Angaben zum Zeitraum der Datensammlung. Dieser sollte präzise dokumentiert werden, wozu strenggenommen auch Angaben zu etwaigen Serverausfällen und Datenverlusten gehören. Wünschenswert wäre außerdem, wenn die aktuellen Informationen zur Social-Media-Plattform selbst festgehalten werden, beispielsweise in Form von Angaben zu Versionen (falls verfügbar) oder zu Nutzerzahlen und zur Nutzerzusammensetzung zum Zeitpunkt der Datenerhebung. In der Praxis ist dies nicht immer einfach. Zusätzlich empfiehlt es sich für Forschende, Screenshots der Nutzeroberfläche einer Social-Media-Plattform zum Zeitpunkt der Datensammlung anzufertigen, da sich diese stetig weiterentwickelt und verändert (Bruns/Weller 2016: 186f). Hiermit ist später zumindest für den Eigengebrauch leichter nachzuvollziehen, über welche Funktionen die Plattform zum jeweiligen Zeitpunkt verfügte. Eine Zusammenstellung von Fragen, welche die Dokumentation der eigenen Social Media-Datensammlung im Forschungskontext vorbereiten kann, findet sich in Schaukasten 11.2 (vgl. dazu auch die Schaukästen zur Planung des Forschungsdatenmanagements in Kapitel 3). Ein Beispiel für eine durchgeführte Datensammlung bei Twitter und deren anschließende Archivierung liefern Kinder-Kurlanda et al. (2017).

Schaukasten 11.2: Leitfragen zur Vorbereitung einer Dokumentation der Social-Media-Datensammlung

- Eigenschaften der Plattform, auf der Daten gesammelt wurden, und ihrer Nutzenden:
  - Aus welchen Social-Media-Plattformen wurden Daten gesammelt?
  - Bezieht sich die Datensammlung auf eine identifizierbare Version? Ggf. können mit Hilfe von Screenshots das Aussehen und die Funktionalitäten der Plattform zum Erhebungszeitpunkt festgehalten werden.
  - Gibt es Informationen zur Nutzerschaft der Plattform(en) zum Erhebungszeitpunkt (z.B. Anzahl der Nutzenden und weitere demographische Informationen)?
  - Sind Datenschutzaspekte und Urheberrechte bei der Datenverarbeitung zu berücksichtigen?
- Beschreibung der erfassten Daten, deren Speicherung und Aufbereitung:
  - Welche Art Daten wurden gesammelt (z.B. Text, Multimedia, Personen, Netzwerke)?
  - In welchem Zeitraum fand die Datensammlung statt?
  - Gibt es Lücken im Zeitraum der Datensammlung (z.B. durch Serverausfälle oder durch Beschränkungen des Datenzugriffs durch API)?
  - Nach welchen Kriterien wurden Daten gesammelt (z.B. Zufallsstichprobe, Suchkriterien wie Stichworte im Text, nach bestimmten Personen etc.)? Ggf. kann eine Abfragesyntax gespeichert werden.
  - Welche Merkmale der gesammelten Daten werden dokumentiert, um sie zu analysieren?
  - Erfolgt die Dokumentation der Daten nach einem strukturierten Schema?
  - Mit welchen Hilfsmitteln wurden die Daten gesammelt? Im Falle von eigens programmierten Tools und Abfragen: Kann hierzu ggf. ein eigener Programmiercode bereitgestellt werden? Im Falle von Drittanbieter-Diensten: Welche Version wurde verwendet, kann diese eindeutig referenziert werden?
  - Wo und wie werden die gesammelten Daten gespeichert (Ort, Dateiformat, -namen)?
  - Wie werden die Originaldaten vor Verlust oder Veränderung im Projekt geschützt?
  - Nach welchen Konventionen werden Arbeitskopien für die Bearbeitung erstellt und gespeichert?
  - Wurden die gesammelten Daten bereinigt (z.B. Dublettenentfernung, Stoppwortentfernung)? Wurden Maßnahmen, Kriterien und Regeln der Bereinigung dokumentiert? Kann hierzu ggf. ein eigener Programmiercode bereitgestellt werden?
  - Wurden die Daten mit zusätzlichen Informationen angereichert (z.B. manuelle oder automatische inhaltliche Codierung, Geo-Informationen)?

Quelle: Eigene Darstellung

11.5 Datenverfall

Trotz erster Bemühungen im Bereich der Archivierung von Social-Media-Datensätzen drohen weitere ernstzunehmende Einbußen der Datenqualität durch die *Flüchtigkeit* des Forschungsgegenstands. Social-Media-Plattformen sind ständigen Dynamiken unterworfen, was nicht selten dazu führt, dass eine Plattform sich vom Zeitpunkt der Datensammlung bis zur Publikation der Forschungsergebnisse ganz wesentlich verändern kann, wie es Karpf (2012: 642ff) am Beispiel von Blogs näher beschreibt. Social-Media-Daten sind zudem in hohem Maße von *Datenverfall* bedroht, und das gleich auf verschiedene Weise.

Innerhalb der Social-Media-Plattformen können Nutzer/innen Informationen erstellen, z.B. Textbeiträge verfassen, Fotos hochladen, Profilseiten gestalten etc. Sie können Inhalte aber auch verändern oder ganz löschen. So können einzelne Posts bei Twitter oder Facebook gelöscht, Freundschaftsverbindungen gekappt oder Nutzerkonten vollständig gelöscht werden. Dies kann innerhalb des Zeitraums der Datensammlung passieren, aber auch zu einem späteren Zeitpunkt.

Wie wir eben gesehen haben, dürfen bei archivierten Twitter-Datensätzen nur die Tweet-IDs geteilt werden. Wer den Datensatz nachnutzen möchte, muss basierend auf den IDs die eigentlichen Tweets noch einmal von Twitter abrufen, ein als *(re-)hydration* bezeichneter Vorgang (Summers 2015). Wurden einzelne Tweets allerdings zwischenzeitlich von ihren Autor/innen oder von Twitter selbst gelöscht, so sind diese auch nicht mehr abrufbar. Von Twitter ist dies durchaus so gewünscht, da es den Nutzenden mehr Selbstbestimmung und Kontrolle über ihre Daten einräumt. Für Forschungskontexte entsteht damit aber ein Dilemma. Bereits kurze Zeit nach der Datensammlung kann es sein, dass ein signifikanter Anteil der Tweets nicht mehr verfügbar ist. Ein archivierter Twitter-Datensatz lässt sich somit wahrscheinlich anhand der Tweet-IDs nicht wieder komplett rekonstruieren.

Bisher gibt es wenig belastbare Informationen, die das genaue Ausmaß dieses Problems beziffern können. Summers (2015) versuchte sich an einem Testlauf und konnte eineinhalb Monate nach der Datensammlung zwischen sieben und zehn Prozent eines großen Twitter-Datensatzes nicht mehr abrufen, da diese zwischenzeitlich gelöscht worden waren.

Wenn nicht die Tweets, sondern die Useraccounts gelöscht werden, entstehen ebenfalls Probleme. Dadurch ist manchmal nachträglich nicht mehr nachvollziehbar, von wem ein konkreter Tweet stammt. Nutzer/innen können aber auch zwischen dem Zeitpunkt einer ursprünglichen Datensammlung und dem Zeitpunkt der Datennachnutzung ihre Profilinformationen geändert haben, wodurch Daten u.U. nicht mehr vergleichbar sind. Bei Twitter kann zudem ein Nutzername (in der Form *@username*) nach der Löschung eines Useraccounts von anderen Nutzenden übernommen werden. Wenn man daher mit den reinen Nutzernamen und nicht mit Nutzer-IDs arbeitet, kann dies zu enormen Verwirrungen führen (Littman 2017b).

Schließlich ist ein großes Problem, dass derzeit nicht dokumentiert wird, wie sich die Social-Media-Plattformen selbst weiterentwickeln. Man weiß daher nicht, wie Twitter oder Facebook zum Zeitpunkt, als ein bestimmter (archivierter) Datensatz entstanden ist, genau aussahen, welche Funktionen den Nutzenden zur Verfügung standen und wie diese genutzt wurden. Hier sind in erster Linie professionelle Gedächtnisinstitutionen gefragt, die die allgemeine Entwicklung von Social-Media-Plattformen als besonderes Kulturgut dokumentieren sollten.

11.6 Fazit

In diesem Beitrag standen die Herausforderungen im Umgang mit Social-Media-Plattformen und anderen Internetdaten als neue Art von sozialwissenschaftlichen Forschungsdaten im Vordergrund. Für jemanden, der in die Arbeit mit Social-Media-Daten einsteigen möchte, ist es wichtig, sich vorab über diese Schwierigkeiten im Klaren zu sein, um nicht im Laufe eines Forschungsprojektes plötzlich festzustellen, dass geplante Vorhaben nicht umsetzbar sind. Dennoch sollen die zahlreichen Problembereiche nicht entmutigen. Es ist durchaus möglich, mit Social-Media-Daten erfolgreiche Forschungsarbeiten durchzuführen. Und auch in den verschiedenen diskutierten Problembereichen werden nach und nach Lösungen erarbeitet. Man sollte sich vor Augen führen, dass es sich hier im Vergleich zur Umfrageforschung und -methodik um einen deutlich jüngeren Bereich handelt, in dem Methoden und Standards noch von der Forschungsgemeinschaft entwickelt und etabliert werden müssen.

Literaturverzeichnis

Blank, Grant (2017): The Digital Divide Among Twitter Users and Its Implications for Social Research. In: Social Science Computer Review 35, 6, S. 679-697. https://doi.org/10.1177/0894439316671698 [Zugriff: 01.06.2018].

Blank, Grant/Lutz, Christoph (2017): Representativeness of Social Media in Great Britain. Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. In: American Behavioral Scientist 61, 7, S. 741-756. https://doi.org/10.1177/0002764217717559 [Zugriff: 01.06.2018].

Boyd, Danah/Crawford, Kate (2012): Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly phenomenon. In: Information, Communication & Society 15, 5, S. 662-679. https://doi.org/10.1080/1369118X.2012.678878 [Zugriff: 01.06.2018].

Bruns, Axel/Weller, Katrin (2016): Twitter as a First Draft of the Present. And the Challenges of Preserving it for the Future. In: Proceedings of the 8th ACM Conference on Web Science (WebSci 16). New York: ACM Press, S. 183-189. https://doi.org/10.1145/2908131.2908174 [Zugriff: 01.06.2018].

Bruns, Axel/Stieglitz, Stefan (2014): Twitter Data: What do they Represent? In: It – Information Technology 56, 5, S. 240-245. https://doi.org/10.1515/itit-2014-1049 [Zugriff: 01.06.2018].

Bruns, Axel/Liang, Yuxian E. (2012): Tools and Methods for Capturing Twitter Data During Natural Disasters. In: First Monday 17, 4. https://doi.org/10.5210/fm.v17i4.3937 [Zugriff: 01.06.2018].

Bruns, Axel (2011): Switching from Twapperkeeper to YourTwapperkeeper. http://mappingonlinepublics.net/2011/06/21/switching-from-twapperkeeper-to-yourtwapperkeeper/ [Zugriff: 01.06.2018].

Burnap, Peter/Rana, Omer/Williams, Matthew/Housley, William/Edwards, Adam/Morgan, Jeffrey/Sloan, Luke/Conejero, Javier (2015): COSMOS. Towards an Integrated and Scalable Service for Analyzing Social Media on Demand. In: International Journal of Parallel, Emergent and Distributed Systems 30, 2, S. 80-100. https://doi.org/10.1080/17445760.2014.902057 [Zugriff: 01.06.2018].

Cha, Meeyoung/Haddadi, Hamed/Benevenuto, Fabricio/Gummadi, Krishna P. (2010): Measuring User Influence in Twitter. The Million Follower Fallacy. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), S. 10-17. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1538 [Zugriff: 01.06.2018].

Choi, Hyonyoung/Varian, Hal (2012): Predicting the Present with Google Trends. In: Economic Record 88, S. 2-9. https://doi.org/10.1111/j.1475-4932.2012.00809.x [Zugriff: 01.06.2018].

Edelman, Benjamin G./Luca, Michael (2014): Digital Discrimination. The Case of Airbnb.com. In: Harvard Business School Working Paper 14-054. http://dx.doi.org/10.2139/ssrn.2377353 [Zugriff: 01.06.2018].

Gaffney, Devin/Matias, J. Nathan (2018): Caveat Emptor, Computational Social Scientists: Large-Scale Missing Data in a Widely-Published Reddit Corpus. https://arxiv.org/abs/1803.05046 [Zugriff: 01.06.2018].

Gaffney, Devin/Puschmann, Cornelius (2014): Data Collection on Twitter. In: Weller, Katrin/Bruns, Axel/Burgess, Jean/Mahrt, Merja/Puschmann, Cornelius (Hrsg): Twitter and Society. New York: Peter Lang, S. 55-68.

Hemphill, Libby/Leonard, Susan H./Hedstrom, Margaret (2018): Developing a Social Media Archive at ICPSR. In: Proceedings of Web Archiving and Digital Libraries (WADL'18). https://deepblue.lib.umich.edu/bitstream/2027.42/143185/1/Developing SOMAR at ICPSR.pdf [Zugriff: 01.06.2018].

Jungherr, Andreas/Schoen, Harald/Jürgens, Pascal (2016): The Mediation of Politics through Twitter. An Analysis of Messages Posted During the Campaign for the German Federal Election 2013. In: Journal of Computer-Mediated Communication 21, 1, S. 50-68. https://doi.org/10.1111/jcc4.12143 [Zugriff: 01.06.2018].

Jungherr, Andreas/Jürgens, Pascal (2014): Through a Glass, Darkly. Tactical Support and Symbolic Association in Twitter Messages Commenting on Stuttgart 21. In: Social Science Computer Review 32, 1, S. 74-89. https://doi.org/10.1177/0894439313500022 [Zugriff: 01.06.2018].

Kaczmirek, Lars/Mayr, Philipp (2015): Deutsche Bundestagswahl 2013. Nutzung von Twitter durch Kandidaten. GESIS Data Archive. ZA5973 Datenfile Version 1.0.0. https://doi.org/doi:10.4232/1.12319 [Zugriff: 01.06.2018].

Karimi, Fariba/Wagner, Claudia/Lemmerich, Florian/Jadidi, Mohsen/Strohmaier, Markus (2016): Inferring Gender from Names on the Web. A Comparative Evaluation of Gender Detection Methods. In: Proceedings of the 25th International Conference Companion on World Wide Web, S. 53-54. https://doi.org/10.1145/2872518.2889385 [Zugriff: 01.06.2018].

Karpf, David (2012): Social Science Research Methods in Internet Time. In: Information, Communication & Society 5, 15, S. 639-661. https://doi.org/10.1080/1369118X.2012.665468 [Zugriff: 01.06.2018].

Kinder-Kurlanda, Katharina E./Weller, Katrin/Zenk-Möltgen, Wolfgang/Pfeffer, Jürgen/Morstatter, Fred (2017): Archiving Information from Geotagged Tweets to Promote Reproducibility and Comparability in Social Media Research. In: Big Data & Society 4, 2. https://doi.org/10.1177/2053951717736336 [Zugriff: 01.06.2018].

Kinder-Kurlanda, Katharina E./Weller, Katrin (2014): "I always feel it must be great to be a Hacker!". The Role of Interdisciplinary Work in Social Media Research. In: Proceedings of the 2014 ACM Web Science Conference WebSci'14. Bloomington, IN, USA. 23.-26. Juni 2014. New York: ACM, S. 91-98. http://dx.doi.org/10.1145/2615569.2615685 [Zugriff: 01.06.2018].

Kitchin, Rob/McArdle, Gavin (2016): What makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. In: Big Data & Society 3, 1. https://doi.org/10.1177/2053951716631130 [Zugriff: 01.06.2018].

Koch, Wolfgang/Frees, Beate (2017): ARD/ZDF-Onlinestudie 2017. Neun von zehn Deutschen online. In: Media Perspektiven 2017, 9, S. 434-446. http://www.ard-zdf-onlinestudie.de/files/2017/Artikel/917_Koch_Frees.pdf [Zugriff: 01.06.2018].

Lazer, David/Radford, Jason (2017): Data ex Machina. Introduction to Big Data. In: Annual Review of Sociology 43, 1, S. 19-39. https://doi.org/10.1146/annurev-soc-060116-053457 [Zugriff: 01.06.2018].

Lazer, David/Kennedy, Ryan/King, Gary/Vespignani, Alessandro (2014): The Parable of Google Flu. Traps in Big Data Analysis. In: Science 343, 6176, S. 1203-1205. https://doi.org/10.1126/science.1248506 [Zugriff: 01.06.2018].

Littman, Justin (2017a): Where to get Twitter Data for Academic Research. https://gwu-libraries.github.io/sfm-ui/posts/2017-11-04-digital-registry [Zugriff: 01.06.2018].

Littman, Justin (2017b): Suspended U.S. Government Twitter Accounts. https://gwu-libraries.github.io/sfm-ui/posts/2017-11-04-digital-registry [Zugriff: 01.06.2018].

Markham, Anette/Buchanan, Elizabeth (2012): Ethical Decision-making and Internet Research 2.0. Recommendations from the AoIR Ethics Working Committee. http://www.aoir.org/reports/ethics2.pdf [Zugriff: 01.06.2018].

McLemee, Scott (2015): The archive is closed. In: Inside Higher Ed. https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research [Zugriff: 01.06.2018].

Metaxas, Panagiotis Takis/Mustafaraj, Eni/Gayo-Avello, Daniel (2011): How (Not) to Predict Elections. In: IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom). IEEE, S. 165-171. https://doi.org/10.1109/PASSAT/SocialCom.2011.98 [Zugriff: 01.06.2018].

Morstatter, Fred/Pfeffer, Jürgen/ Liu, Huan (2014): When is it Biased? Assessing the Representativeness of Twitter's Streaming API. In: Proceedings of Web ScienceTrack at the 23rd Conference on the WWW. New York: ACM, S. 555-556. https://doi.org/10.1145/2567948.2576952 [Zugriff: 01.06.2018].

Morstatter, Fred/Pfeffer, Jürgen/Liu, Huan/Carley, Kathleen M. (2013): Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, S. 400-408. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071/6379 [Zugriff: 01.06.2018].

Nishioka, Chifumi/Scherp, Ansgar/Dellschaft, Klaas (2015): Manual Tweet Classification. GESIS Data Archive. http://dx.doi.org/10.7802/82 [Zugriff: 01.06.2018].

Olteanu, Alexandra/Castillo, Carlos/Diaz, Fernando/Vieweg, Sarah (2014): CrisisLex. A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In: Proceedings of 8th International AAAI Conference on

Weblogs and Social Media (ICWSM'14), Ann Arbor, US. Juni 2014, S. 376-385. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8091 [Zugriff: 01.06.2018].

O'Reilly, Tim (2005): What is Web 2.0? Design Patterns and Business Models for the Next Generation of Software. http://oreilly.com/web2/archive/what-is-web-20.html [Zugriff: 01.06.2018].

Osterberg, Gayle (2017): Update on the Twitter Archive at the Library of Congress. https://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2/ [Zugriff: 01.06.2018].

Pfeffer, Jürgen/Morstatter, Fred (2016): Geotagged Twitter Posts from the United States. A Tweet Collection to Investigate Representativeness. GESIS Data Archive. http://dx.doi.org/10.7802/1166 [Zugriff: 01.06.2018].

Resnick, Paul/Adar, Eytan/Lampe, Cliff (2015): What Social Media Data we are Missing and How to Get it. In: The ANNALS of the American Academy of Political and Social Science 659, 1, S. 192-206. https://doi.org/10.1177/0002716215570006 [Zugriff: 01.06.2018].

Ruths, Derek/Pfeffer, Jürgen (2014): Social Media for Large Studies of Behavior. In: Science 346, 621, S. 1063-1064. https://doi.org/10.1126/science.346.6213.1063 [Zugriff: 01.06.2018].

Ruiz, Carlos/Domingo, David/Mico, Josep L./Diaz-Noci, Javier/Meso, Koldo/ Masip, Pere (2011): Public Sphere 2.0? The Democratic Qualities of Citizen Debates in Online Newspapers. In: The International Journal of Press/Politics 16, 4, S. 463-487. https://doi.org/10.1177/1940161211415849 [Zugriff: 01.06.2018].

Schmidt, Jan (2009): Das neue Netz. Merkmale, Praktiken und Folgen des Web 2.0. Konstanz: UVK.

Schroeder, Ralph (2014): Big Data and the Brave New World of Social Media Research. In: Big Data & Society 1, 2, S. 1-11. https://doi.org/10.1177/2053951714563194 [Zugriff: 01.06.2018].

Sloan, Luke/Quan-Haase, Anabel (2016): The Sage Handbook of Social Media Research Methods. Thousand Oaks, CA: SAGE.

Sloan, Luke/Morgan, Jeffrey (2015): Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. In: PLoS ONE 10, 11, e0142209. https://doi.org/10.1371/journal.pone.0142209 [Zugriff: 01.06.2018].

Song, Melodie YJ./Gruzd, Anatoliy (2017): Examining Sentiments and Popularity of Pro- and Anti-Vaccination Videos on YouTube. In: Proceedings of the 8th International Conference on Social Media & Society. New York: ACM Press. https://doi.org/10.1145/3097286.3097303 [Zugriff: 01.06.2018].

Stone, Biz (2010): Tweet Preservation. https://blog.twitter.com/official/en_us/a/2010/tweet-preservation.html [Zugriff: 01.06.2018].

stuck_in_the_matrix (2015a): I have every publicly available Reddit comment for research: ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment [Zugriff: 01.06.2018].

stuck_in_the_matrix (2015b): Complete Public Reddit Comments Corpus. https://archive.org/details/2015_reddit_comments_corpus [Zugriff: 01.06.2018].

Summers, Ed (2015): Tweets and Deletes. Silences in the Social Media Archive. https://medium.com/on-archivy/tweets-and-deletes-727ed74f84ed#.pay32r3eu [Zugriff: 01.06.2018].

Summers, Ed (2014): Ferguson-tweet-ids. https://archive.org/details/ferguson-tweet-ids [Zugriff: 01.06.2018].

Thomson, Sarah D. (2016): Preserving Social Media. DPC Technology Watch Report. http://dx.doi.org/10.7207/twr16-01 [Zugriff: 01.06.2018].

TREC (2011): Tweets2011. http://trec.nist.gov/data/tweets/ [Zugriff: 01.06.2018].

Tufekci, Zeynep (2014): Big Questions for Social Media Big Data. Representativeness, Validity and Other Methodological Pitfalls. In: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM), S. 505-514. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062 [Zugriff: 01.06.2018].

Weller, Katrin/Kinder-Kurlanda, Katharina E. (2016): A Manifesto for Data Sharing in Social Media Research. In: Proceedings of the 8th ACM Conference on Web Science (WebSci 16). New York: ACM Press, S. 166-172. https://doi.org/10.1145/2908131.2908172 [Zugriff: 01.06.2018].

Weller, Katrin (2015): Accepting the Challenges of Social Media Research. In: Online Information Review 39, 3, S. 281-289. https://doi.org/10.1108/OIR-03-2015-0069 [Zugriff: 01.06.2018].

Weller, Katrin/Kinder-Kurlanda, Katharina E. (2015): Uncovering the Challenges in Collection, Sharing and Documentation. The Hidden Data of Social Media Research. In: Standards and Practices in Large-Scale Social Media Research. Papers from the 2015 ICWSM Workshop. Proceedings Ninth International AAAI Conference on Web and Social Media. Ann Arbor, MI: AAAI Press, S. 28-37. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewFile/10657/10552 [Zugriff: 01.06.2018].

Weller, Katrin (2014): Twitter und Wahlen. Zwischen 140 Zeichen und Milliarden von Tweets. In: Reichert, Ramon (Hrsg.): Big Data. Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie. Bielefeld: transcript, S. 239-257.

Weller, Katrin/Kinder-Kurlanda, Katharina E. (2014): "I love thinking about ethics!" Perspectives on Ethics in Social Media Research. In: Selected Papers of Internet Research (SPIR). Proceedings of ir15 – Boundaries and Intersections, Deagu, South Korea. https://spir.aoir.org/index.php/spir/article/view/997 [Zugriff: 01.06.2018].
Zimmer, Michael/Kinder-Kurlanda, Katharina E. (Hrsg.) (2017): Internet Research Ethics for the Social Age. New Challenges, Cases, and Contexts. New York: Peter Lang.
Zimmer, Michael (2015): The Twitter Archive at the Library of Congress. Challenges for Information Practice and Information Policy. In: First Monday 20, 7. https://doi.org/10.5210/fm.v20i7.5619 [Zugriff: 01.06.2018].
Zimmer, Michael/Proferes, Nicholas J. (2014a): Privacy on Twitter, Twitter on Privacy. In: Weller, Katrin/Bruns, Axel/Burgess, Jean/Mahrt, Merja/Puschmann, Cornelius (Hrsg): Twitter and Society. New York: Peter Lang, S. 169-181.
Zimmer, Michael/Proferes, Nicholas J. (2014b): A topology of Twitter research. Disciplines, Methods, and Ethics. In: Aslib Journal of Information Management 66, 3, S. 250-261. https://doi.org/10.1108/AJIM-09-2013-0083 [Zugriff: 01.06.2018].

Linkverzeichnis

Angela Merkel auf Facebook: https://www.facebook.com/AngelaMerkel/ [Zugriff: 01.06.2018].
CrisisLex: http://crisislex.org/ [Zugriff: 01.06.2018].
DataSift: http://datasift.com [Zugriff: 01.06.2018].
Discover Text: http://discovertext.com [Zugriff: 01.06.2018].
Facebook: http://www.facebook.com [Zugriff: 01.06.2018].
Foursquare: http://foursquare.com [Zugriff: 01.06.2018].
Instagram: http://www.instagram.com [Zugriff: 01.06.2018].
ICWSM – International Conference on Web and Social Media: http://icwsm.org [Zugriff: 01.06.2018].
ICWSM Dataset Sharing Service: http://icwsm.cs.mcgill.ca [Zugriff: 01.06.2018].
Inter-university Consortium for Political and Social Research (ICPSR): https://www.icpsr.umich.edu/icpsrweb/ [Zugriff: 01.06.2018].
Jack Dorseys erster Tweet: https://twitter.com/jack/status/20 [Zugriff: 01.06.2018].
KONECT. The Koblenz Network Collection: http://konect.uni-koblenz.de/ [Zugriff: 01.06.2018].
LinkedIn: http://linkedin.com [Zugriff: 01.06.2018].
Max Planck Institute for Software Systems: The Twitter Project Page at MPI-SWS: http://twitter.mpi-sws.org/ [Zugriff: 01.06.2018].
NVIVO: http://www.qsrinternational.com/nvivo/nvivo-products [Zugriff: 01.06.2018].
Pinterest: http://pinterest.com [Zugriff: 01.06.2018].
reddit: http://reddit.com [Zugriff: 01.06.2018].
Social Feed Manager: https://gwu-libraries.github.io/sfm-ui/ [Zugriff: 01.06.2018].
Social Media & Society Conference: http://socialmediaandsociety.org/ [Zugriff: 01.06.2018].
TREC – Text Retrieval Conference: https://trec.nist.gov/ [Zugriff: 01.06.2018].
Tweet Archivist: http://tweetarchivist.com [Zugriff: 01.06.2018].
Twitter: http://twitter.com [Zugriff: 01.06.2018].
Tumblr: http://tumblr.com [Zugriff: 01.06.2018].
Vkontakte: http://vk.com [Zugriff: 01.06.2018].
Web Science Conference: http://www.webscience.org/category/acm-websci/ [Zugriff: 01.06.2018].
Wikipedia: http://www.wikipedia.org [Zugriff: 01.06.2018].
Wikipedia Dumps: Wikipedia:Database_download: https://en.wikipedia.org/wiki/Wikipedia:Database_download [Zugriff: 01.06.2018].