Im Frühjahr 2023 starteten die Deutsche Nationalbibliothek (DNB) und das Science Data Center for Literature eine Archivierunginitiative für das deutschsprachige Twitter per Crowdsourcing. Im Beitrag werden die Motivation für die Initiative, die technische und organisatorische Vorgehensweise und schließlich die Ergebnisse beschrieben und ein Ausblick für die künftige Bereitstellung und Nutzung der Daten gegeben.

In spring 2023, the German National Library and the Science Data Center for Literature launched a crowdsourced archiving initiative for German-language Tweets. The article describes the motivation behind the initiative, the technical and organisational aspects involved and finally the results. It also provides an outlook on the provision and use of the data in the future.

CLAUS-MICHAEL SCHLESINGER, BRITTA WOLDERING

»Den Heuhaufen archivieren«: Archivierungsinitiative für deutschsprachige Tweets

Einleitung

Social Media ist Teil des kulturellen Erbes, und mit jeder Plattform, die entsteht, aufblüht und wieder zugrunde geht oder sich weitgehend transformiert, verschwinden Teile der damit verbundenen Gesellschafts- und Kulturgeschichte.[1] Denn während zum Beispiel für gedruckte Publikationen gut ausdifferenzierte Abläufe definiert sind, die dafür sorgen, dass Druckerzeugnisse weitgehend erhalten werden, existieren für Social-Media-Publikationen keine vergleichbaren Grundlagen. Für Twitter wurde dieser Zustand durch die weitreichende Transformation der Plattform nach der Übernahme durch ein Investorenkonsortium um Elon Musk im Herbst 2022 deutlich. Dabei rückt Twitter bereits seit einigen Jahren auch als historischer Gegenstand in den Blick, das heißt als Medium historischer Quellen und als Medium mit eigener Geschichte.[2] Eine umfassende Archivierung findet im deutschsprachigen Raum bisher nur ansatzweise institutionalisiert statt.[3] Sammlungen werden eher von einzelnen Forscher*innen oder von Forschungsprojekten angelegt,[4] können aber aufgrund der restriktiven Terms of Service und aus Gründen des Urheberrechts oft nicht so verfügbar gemacht werden, wie dies für nachvollziehbare und reproduzierbare Forschungsergebnisse notwendig wäre.

Die neuen Zugriffsbeschränkungen von Twitter – eine starke Mengenbegrenzung für den verbliebenen kostenfreien Zugang und kein kostenfreier Zugang mehr für Forschende – stellen ganz unterschiedliche Forschungsdisziplinen vor weitreichende Herausforderungen. Erstens werden selbst kleine themen- oder autorenorientierte Sammlungen dadurch finanziell nicht mehr tragbar oder finanziell zumindest so aufwendig, dass eine Archivierung von zusätzlichen Mittelzuweisungen abhängig und die Bereitschaft zur Archivierung dadurch gesenkt wird. Zweitens fungierte Twitter in den letzten Jahren zunehmend als alleiniger Data Provider für die Forschung. Obwohl die Zugangspolitik der Plattform im Vergleich mit anderen großen Plattformen großzügig gestaltet war, verbieten die Terms of Service eine Weitergabe von Datensätzen, auch für die wissenschaftliche Nachnutzung. Üblicherweise werden daher Tweet-Datensätze als Listen von Tweet-IDs veröffentlicht. Mithilfe der Tweet-IDs können dann unter Nutzung der Twitter-APIs vollständige Datensätze rekonstruiert werden. Neben der Einschränkung einer möglichen Zirkulation von Daten führt dieses Vorgehen auch dazu, dass bei der Rekonstruktion oder Rehydrierung, so die übliche Bezeichnung, zwischenzeitlich gelöschte Tweets nicht mehr Teil des Datensatzes sind. Aus Sicht einer auf Reproduzierbarkeit angelegten Forschung ist das schlecht, andererseits behalten auf diese Weise die Nutzer*innen eine gewisse Hoheit über ihre Daten, was durchaus als positiver Aspekt im Umgang mit ethischen Fragen bei der Archivierung von Social-Media-Daten verstanden werden kann. Schwerer wiegt allerdings, dass Twitter als alleiniger Data Provider für Forschung und Archivierung einen *single point of failure* darstellt, der nicht einfach ersetzt werden kann.[5] Auch Kulturerbeeinrichtungen sind von der weitreichenden Monetarisierung des Zugangs zu den Twitter-APIs betroffen, sofern sie für die Archivierung auf diesen

Zugang zurückgreifen. Andere Techniken für die Webarchivierung und Dokumentation von Webseiten, die an die ausgelieferten Seiten der Webanwendung anschließen – also browserbasierte Webarchivierung, Screenshots und ähnliches – sind weiterhin nutzbar, allerdings ebenfalls nur eingeschränkt, weil die Auslieferung von Tweets über die Webanwendung inzwischen ebenfalls zahlenmäßig beschränkt ist.[6]

Die weitreichenden Transformationen von Twitter, die spätestens mit der Umbenennung der Plattform in »X.com« das Soziale Medium Twitter als Zusammenhang von Plattformfunktionen und Interaktionen an sein historisches Ende geführt hat, führen zu einem gesteigerten Bedarf auch institutioneller Erhaltung von Tweets als Teil des digitalen Kulturerbes. Dabei müssen Tweets überhaupt erhalten, das heißt archiviert werden,[7] zudem ist auch an eine institutionell geregelte Archivierung und Bereitstellung zu denken, damit beispielsweise die Herausgeberin der Gesammelten Werke einer Autorin, die heute vielleicht noch unbekannt ist, in einer Tweet-Sammlung die frühen Texte findet.[8]

Die Archivierungsinitiative

Twitter kündigte Ende 2022 an, die Zugangsbedingungen für die bestehenden APIs zu ändern, wobei der Zeitpunkt der Änderung und die neuen Bedingungen lange unklar blieben. Diese Ankündigung war die Motivation für die gemeinsame Initiative der Deutschen Nationalbibliothek (DNB) und des Science Data Center for Literature, möglichst viele deutschsprachige Tweets aus dem Twitter-Archiv bei der DNB zu archivieren und dafür ein Crowdsourcing zu organisieren. Zugleich war durch die Ankündigung ein gewisser Zeitdruck entstanden. Im November 2022 wurde mit den Vorarbeiten für die Archivierungsinitiative begonnen und im Februar 2023 ein Unterstützungsaufruf an verschiedene Forschungscommunities verschickt. Mit dem Abschalten des Academic Research Access und anderer Access-Level am 29. April 2023 kam die Initiative zu ihrem Ende. Im Folgenden wird beschrieben, wie die Initiative technisch und organisatorisch angelegt war.

Archivierung von Tweets

Die institutionelle Archivierung von Tweets an Bibliotheken und Archiven ist im deutschsprachigen Raum über die jeweiligen Sammlungsziele und -richtlinien oft themen- oder accountgebunden. Dabei sind auch themen- oder accountgebundene Sammlungen auf die Klärung sammlungsstrategischer, technischer und rechtlicher Fragen angewiesen, sodass anlässlich entsprechender Sammlungen bereits Grundlagen für die Archivierung von Tweets durch Archive und Bibliotheken gelegt sind.[9] Außerdem gab und gibt es international weitreichende Anstrengungen zur institutionellen Archivierung von Tweets.[10] So hat die Library of Congress (LoC) bis zum Jahr 2017 in Zusammenarbeit mit der Plattform ein vollständiges Archiv sämtlicher Tweets angelegt, das den Zeitraum von 2006, dem Beginn der Plattform, bis einschließlich 2017 umfasst. Mit Beginn des Jahres 2018 wurde die vollständige Sammlung zugunsten einer auf Auswahl gerichteten Sammlung aufgegeben. Das Tweet-Archiv der LoC ist ein *dark archive*, die letzten verbindlichen Aussagen dazu finden sich nach unserem Wissen in einem Blogpost, in dem das Ende der vollständigen Sammlung mitgeteilt und begründet wird.[11] Andere Nationalbibliotheken wie die Bibliothèque nationale de France oder die neuseeländische Nationalbibliothek setzen auf ereignisorientierte, thematische und accountorientierte Auswahlregeln.[12]

Auch in der Forschung überwiegen themen- und accountorientierte Sammlungen, weil Datensätze in der Regel anlässlich von Forschungsfragen zusammengestellt werden. Dabei ist die Zusammenstellung eines Korpus für ein bestimmtes Themenfeld ausgehend von einer Forschungsfrage oft weitreichender nutzbar, etwa wenn thematische Sammlungen auch für die Bearbeitung anderer Fragen auf dem gleichen Feld verwendet werden können.[13] Nicht themengebundene, auf deutschsprachige Tweets orientierte Sammlungen finden sich vorwiegend in den Bereichen Linguistik und Sozialwissenschaften.[14] Darüber hinaus finden sich weitere umfangreiche, weder themen- noch sprachgebundene Sammlungen mit einem Anspruch auf Vollständigkeit oder repräsentative Auswahl, die in internationalen Forschungskontexten erstellt und an Forschungseinrichtungen verwaltet und bereitgestellt werden.[15] Forschung zu Twitter basiert dabei nicht notwendig auf Tweets. Insbesondere in der Sozialforschung spielen Netzwerkdaten eine wichtige Rolle. Bestimmte Netzwerkstrukturen können von Tweet-Texten und -Metadaten abgeleitet werden, beispielsweise über zitierte Tweets, Retweets oder die Nennung von Accountnamen in einem Tweet. Insbesondere für Friends-Follower-Netzwerke, die eine direkte Verbindung zwischen Accounts über die ›Folgen‹-Funktion anzeigen, sind aber accountbezogene Metadaten notwendig, die in der Regel nicht Teil von Tweet-Sammlungen sind, weil sie nur über eine eigene mengenbegrenzte API verfügbar sind bzw. waren und daher separat abgerufen werden mussten.[16]

Gegenstand der Archivierungsinitiative

Die Archivierung von Tweets und zugehörigen Metadaten setzt, wie alle Archivierungsbemühungen, ein spezifisches Verständnis des Gegenstands voraus. Die zu erhaltenden Eigenschaften müssen mit erwarteten Umgangsformen und Erkenntnisinteressen abgestimmt und mit Blick auf Archivierbarkeit bewertet und dokumentiert werden. Das elementare Objekt der Sammlung ist ein Tweet, die Sammlung ist also eine Menge von einzelnen Tweets. Durch die Metadaten sind Informationen zu bestimmten Eigenschaften und Kontexten eines Tweets erhalten, darunter der sendende Account, ob

der Tweet eine Antwort auf einen anderen Tweet ist und wenn ja auf welchen, die Anzahl der Likes und Retweets zum Zeitpunkt des Abrufs usw. Ziel der Archivierungsinitiative war die umfassendste Dokumentation aller Tweets durch den Abruf möglichst umfangreicher Metadaten. Die Entscheidung für möglichst umfangreiche Metadaten ist durch den breit streuenden webarchivarischen Fokus begründet, anders als etwa bei Sammlungen, die für konkrete Forschungsfragen erstellt werden. Kathrin Passig hat in einem für uns wegweisenden Vortrag auf der Jahreskonferenz des DHd-Verbands vorgerechnet, wie umfangreich ein Archiv mit allen deutschsprachigen Tweets ist, verbunden mit der impliziten Forderung, das möglichst bald in die Wege zu leiten.[17] Hintergrund für diese Forderung ist das von Passig in einem anderen Zusammenhang präsentierte Bild von im Internet veröffentlichten Texten als Nadel im Heuhaufen, die wegen der kurzen Halbwertszeit von Texten im Internet zusammen mit dem Heu schneller verschwinden als sie gefunden oder als Nadeln überhaupt identifiziert werden können. Ein Beispiel dafür sind Texte von Autor*innen, deren Werk erst zu einem späteren Zeitpunkt für vollständig erhaltenswert erachtet wird, wenn Texte im Internet, zum Beispiel Tweets, längst verschwunden oder hinter für Herausgeber*innen undurchdringlichen Paywalls versperrt sind.[18] In diesem Bild ist jeder Tweet ein Text oder eine eigenständige Publikation, das Einzelobjekt durch die zugehörigen Metadaten identifiziert. Die Nadeln sind dabei später nur zu finden, wenn der ganze Heuhaufen archiviert wird. Tweets sind aus dieser Perspektive historische Texte, die erhaltenswert sind, weil sie möglicherweise einmal wichtig werden könnten. Gleichzeitig impliziert das Bild des Heuhaufens auch eine Menge Heu, das heißt sehr viele Tweets. In den Blick rücken damit korpusorientierte Fragestellungen, also Forschungen, die sich nicht mit den Nadeln, sondern mit dem Heu beschäftigen und etwa textstatistische oder auf aggregierte Metadaten gerichtete Fragestellungen bearbeiten. Der Heuhaufen als Menge von Tweets ermöglicht damit alles, was zwischen den Grenzfällen einer Suche nach einzelnen Tweets einer Autorin oder eines Autors und korpusorientierten Analysen mit großen Datenmengen liegt.

Tweets bestehen nicht allein aus Texten und Metadaten. Sie erscheinen zunächst im Kontext der Plattform in accountbezogenen Timelines, Conversations oder Suchergebnislisten. Neben den Tweet-Daten rücken hier auch visuelle und funktionale Aspekte in den Blick. Für ein Verständnis von Social Media aus mediengeschichtlicher Sicht sind diese Aspekte grundlegend relevant. Bruns und Weller betonen daher die Notwendigkeit, im Umgang mit Social-Media-Plattformen auch solche Aspekte zu erhalten oder zu dokumentieren.[19] Dokumentation kann dabei als Aufgabe verstanden werden, die Forschung und Archivierung gleichermaßen betrifft. Denn erstens ist für ein wissenschaftliches Verständnis historischer Daten auch ein Verständnis der historischen Plattform als primärem Kontext relevant, zweitens ist Twitter als Plattform aus webarchivarischer Sicht ein grenzenloses Objekt (*boundless object*), das nicht vollständig erhalten werden kann.[20] Plattformfunktionen, -gestaltung und -nutzung können aber dokumentiert, die veröffentlichten Texte mit Metadaten archiviert werden. Allerdings geht durch die Archivierung der dynamische Kontext der Plattform verloren. Es ist also zu unterscheiden zwischen Tweets auf der Plattform und archivierten Tweets, die mit ihrer Archivierung von der Plattform getrennt und zu einem auch eigenständigen Gegenstand oder Archivobjekt werden.

Die hier dokumentierte Sammlung deutschsprachiger Tweets war aus Zeitgründen auf die Archivierung via Twitter-Search-API beschränkt. Als elementarer Gegenstand wurden »deutschsprachige Tweets« bestimmt, wobei der Fokus auf der Archivierung von originalen Tweets lag. Retweets wurden im verwendeten Suchstring aus Zeit- und Kapazitätsgründen ausgeschlossen. Medien (Bilder, Video, Audio) werden nicht über die Twitter-API ausgespielt und wurden daher nicht gespeichert. Die entsprechenden URIs zu den Mediendateien sind Teil der gespeicherten Metadaten. Friends-Follower-Daten wurden aufgrund der Fokussierung von Tweets als primärem Gegenstand und aus Kapazitätsgründen ebenfalls nicht gespeichert. Mit der Studie von Hammer zur deutschsprachigen Twittersphäre liegt eine entsprechende Netzwerkanalyse auf Grundlage einer umfangreichen Datenerhebung bereits vor.[21] Die Beschränkung auf deutschsprachige Tweets ist durch den Sammlungsauftrag der DNB und darüber hinaus pragmatisch begründet. Die DNB hat den gesetzlichen Auftrag, alles zu sammeln, zu erschließen, zu archivieren und zugänglich zu machen, was seit 1913 in oder über Deutschland erscheint. Seit 2006 umfasst der Sammlungsauftrag auch die sogenannten unkörperlichen Medienwerke, wozu auch Websites und in diesem Zuge auch Inhalte von Social-Media-Plattformen gehören. Social Media sind Teil der hybriden Medienlandschaft Deutschlands und eine wichtige Quelle der Kultur- und Gesellschaftsgeschichte. Das deutschsprachige oder auf Deutschland bezogene Twitter gehört also grundsätzlich in den Sammlungsauftrag der DNB.

Bei der Operationalisierung wurde die sprachbezogene Auswahl von Tweets in Anbetracht der geforderten Eile als bester Kompromiss zwischen notwendiger Begrenzung und gewünschter Breite der Auswahl identifiziert. Wünschenswert und für den Fall ausreichender Kapazitäten optional vorgesehen war, mit zusätzlichen Verfahren auch multilinguale Tweets im Sinne des Sammlungsauftrags zu finden und ergänzend zu archivieren, etwa durch eine nachgeordnete Identifikation von Accounts oder Hashtags auf Basis der sprachgebundenen Sammlung. Ein Beispiel dafür sind englischsprachige Tweets deutscher Politiker*innen. Mit der institutionellen Archivierung an der DNB (im Unterschied zur

Sammlung durch einzelne Forscher*innen oder Forschungsgruppen) verbunden ist, dass die Sammlung später nach den gesetzlich geregelten und institutionell etablierten Zugangsregeln der DNB für rechtlich geschütztes Material bereitgestellt und genutzt werden kann.

Methodik

Die Sammlung wurde unter Nutzung der Twitter-Search-API mit dem Zugangsmodus »Academic Research Access« umgesetzt. Für die sprachliche Einschränkung wurde das von Twitter vergebene Sprachlabel verwendet. Zielvorgabe für den Suchstring war deshalb: alle Tweets, die von Twitter als deutschsprachig gelabelt sind. Die Search-API erforderte zum Zeitpunkt der Sammlung mindestens eine eigenständige Zeichenkette als Suchstring, das heißt eine Suche mit Festlegung des Sprachlabels allein (als optionalem Attribut) war nicht möglich. Die Sammlungen von Scheffler[22] und X-GTA[23] arbeiten mit einem Most-Frequent-Words-Ansatz unter besonderer Berücksichtigung von Verbindungswörtern, um möglichst umfangreich und dabei trennscharf deutschsprachige Tweets zu identifizieren. Vorteil dieser Strategie ist, dass sie prinzipiell ohne Anwendung der Spracherkennung von Twitter funktioniert, dabei allerdings sehr kurze Tweets oder Tweets, die lediglich Hashtags, Medien oder Emojis enthalten, nicht mit abgedeckt sind. Auch aus Zeitgründen wurde für die Sammlung schließlich ein Suchstring definiert, der die erforderliche Zeichenkette als negatives Kriterium enthält. Das bedeutet, dass die Suchergebnisse die gesetzte Zeichenkette *nicht* enthalten.[24] Der Suchstring wird damit recht überschaubar: »lang:de -krzlfhrrrrbnldgh -is:retweet«. Die Minuszeichen stehen dabei für ein Ausschlusskriterium. Gesucht und gesammelt werden damit also alle Tweets, die von Twitter als deutschsprachig gelabelt sind, die nicht die Zeichenkette »krzlfhrrrrbnldgh« enthalten und die kein Retweet sind. Retweets sind Tweets, die lediglich erneut gepostet werden. Diese Eigenschaft ist für Analysen wünschenswert, erhöht aber die Anzahl der herunterzuladenden Tweets und damit den Bedarf an entsprechender Zugangs-Quota für die Search-API beträchtlich. Für die Sammlung wurden daher originale Tweets (Ausgangstweets und Antworten) priorisiert. Vor Einsatz des negativen Suchworts wurde eine Suche mit dem Suchwort als positivem Suchkriterium durchgeführt, um sicherzustellen, dass keine Tweets existieren, die die Zeichenkette beinhalten.

Die Twitter-eigene Sprachklassifizierung ist in einem Blogpost der Entwickler*innen teilweise dokumentiert.[25] Eine unabhängige Prüfung und Qualitätskontrolle für die Klassifizierung deutschsprachiger Tweets liegt unseres Wissens nach bisher nicht vor.[26] Aus diesem Grund wurden im Zuge der Entwicklung des Suchstrings Stichproben mit jeweils hundert Tweets aus den Jahren 2008–2022 mit einem jeweils zweijährigen Abstand manuell annotiert. Die Stichproben beinhalten jeweils 10–15 % *false positives*, d. h. nicht deutschsprachige Tweets, die als deutschsprachig gelabelt sind. *False negatives*, also deutschsprachige Tweets, die nicht als deutschsprachige gelabelt sind, sind nur mittels umfangreicher Daten sinnvoll zu ermitteln. Eine entsprechende Untersuchung steht nach unserem Wissen noch aus, ist aber für die Einschätzung von Sammlungen und davon abgeleiteten Geltungsansprüchen in Bezug auf Vollständigkeit ein Desiderat.

Mit dem Zugangslevel »Academic Research Access« erlaubte Twitter zum Zeitpunkt der Sammlung einen Zugang zum gesamten Archiv. Die Nutzung der Twitter-Search-API mit Academic Research Access bietet eine sehr hohe Zuverlässigkeit, die archivarischen Ansprüchen vollkommen genügt.[27] Die Mengenbeschränkung für Accounts mit Academic Research Access betrug in der Regel zehn Millionen Tweets pro Monat. Academic Research Access war zum Zeitpunkt der Sammlung nach einer Beantragung im Developer-Portal von Twitter nach Nachweis und Überprüfung der geplanten wissenschaftlichen Nutzung verfügbar. Im Laufe des Sammlungszeitraums wurden neu gestellte Anträge allerdings nicht mehr bearbeitet. Die zu Beginn der Sammlung gültigen Zugänge für entsprechend freigeschaltete Entwickler*innen-Accounts wurden schließlich im Zuge der Umstellung auf ein neues Zugriffsmodell und die umfangreiche Monetarisierung des Zugangs bis Ende April 2023 sukzessive deaktiviert.

Die Anzahl der Tweets für einen bestimmten Suchstring wurde vorbereitend mithilfe der Twitter-Count-API ermittelt. Diese Schnittstelle lieferte auf eine Suchanfrage, wie sie auch an die Search-API gestellt werden konnte, lediglich die Anzahl der Tweets für die Anfrage zurück. Für die verwendete Suchanfrage wurde auf diese Weise ein Umfang von rund drei Milliarden Tweets ermittelt. Die Monetarisierung des Zugangs war beim Start der Initiative bereits absehbar, die ermittelte Menge von Tweets war also lediglich durch parallelen Zugriff mit mehreren Accounts denkbar. Mit einem Aufruf wurden daher Forscher*innen zur Unterstützung der Initiative eingeladen. Zur Umsetzung dieses Crowdsourcing-Ansatzes wurden die mittels Count-API ermittelten erwarteten Tweets auf mehrere Batches aufgeteilt, die jeweils durch ein Start- und ein Enddatum definiert sind und eine Million Tweets umfassen. Für die Koordination der gemeinsamen Archivierung wurde eine Webanwendung entwickelt und auf der DNB-Infrastruktur bereitgestellt, die eine Reservierung einzelner Batches mit nachfolgender Lieferung oder das Auslösen der Archivierung reservierter Batches auf dem DNB-Server ermöglichte.[28] Auf diese Weise konnten Teilnehmer*innen in einer Sitzung jeweils 1–10 Batches reservieren und archivieren. Die Archivierung der einzelnen Batches wurde mit dem Python-Paket Twarc realisiert.[29] Twarc wurde von der Initiative Documenting the Now mit einem Fokus auf

die Archivierung von Tweets entwickelt. Es ermöglicht den einfachen Abruf von umfangreichen Metadaten über die API und schreibt Prozessmetadaten für jeden einzelnen Tweet in die Ergebnisdaten, insbesondere den Zeitpunkt des Zugriffs auf die API, den verwendeten Suchstring und die verwendete Programmversion.

Die für eine vollständige Archivierung der rund drei Milliarden deutschsprachigen Tweets notwendige Dauer hängt bei gleichbleibender Quota von der Anzahl der parallel durchführbaren Archivierungen vordefinierter Batches und damit von der Anzahl der beteiligten Accounts ab. Eine Rechnung ergab für die Durchführung mit einem einzigen Account eine Dauer von 30 Jahren, mit zehn Accounts drei Jahre, mit fünfzig Accounts sieben Monate und mit 350 Accounts weniger als einen Monat. Die Anzahl der Teilnehmer*innen betrug schließlich zwischen 20 und 30 Personen. Die aktive Teilnahme an der Aktion erfolgte mithilfe von Zugangsdaten für die Webanwendung. Die Zugänge wurden aus Datenschutzgründen nicht gespeichert. Die archivierten Daten sind generisch gespeichert, das heißt accountbezogene oder anderweitige persönliche Informationen der Teilnehmer*innen wurden nicht gespeichert.

Ergebnisse und Ausblick

Die Sammlung wurde mit den frühen Tweets begonnen und dann chronologisch fortgesetzt. Während der Sammlungsphase von Februar bis April 2023 konnte ein umfangreiches Korpus »Frühes Twitter« erstellt werden, das den Zeitraum vom Beginn der Plattform im Jahr 2006 durchgehend bis einschließlich Juni 2011 umfasst. Das Korpus enthält rund 220 Millionen Tweets von mehr als sechs Millionen Accounts. Forschungsrelevante Metadaten sind bereits strukturiert erfasst und umfassen etwa Hashtags, Hyperlinks, Timestamps, Mentions, die Anzahl Likes, Retweets und Replies für jeden einzelnen Tweet zum Zeitpunkt des Downloads, Conversation-IDs und weitere.[30] Auf Basis der Daten ist die Zusammenstellung von Subkorpora zur Bearbeitung spezifischer Fragestellungen möglich, z. B. Hashtag-Analysen oder accountorientierte Analysen.

Im Zuge der Twitter-Archivieren-Initiative ist der DNB ein Forschungsdatensatz zur Archivierung angeboten worden, der das Korpus »Frühes Twitter« hervorragend ergänzt und von der DNB übernommen wurde. Dieses Korpus enthält ca. 2 Milliarden deutschsprachige Tweets und umfasst den Zeitraum Juli 2014 bis März 2023. Das Korpus wurde unter Zugriff auf die Twitter-Streaming-API mit einem Most-Frequent-Words-Ansatz zusammengestellt.[31] Neben den Tweet-Texten wurden in dem Korpus nur einzelne Metadaten gespeichert, nämlich Tweet- und User-ID, Zeitpunkt des Postings, Reply-to-ID und Geodaten. Im Zuge einer weiteren Erschließung werden Hashtags und Links aus den Tweet-Texten automatisiert ausgelesen und dem Datensatz in strukturierter Form beigefügt. Die Daten liegen als CSV-Dateien vor.

Für die DNB ist die Sammlung, Archivierung und Bereitstellung von Social-Media-Daten in Form strukturierter Rohdaten ein Experimentierfeld. Die Bereitstellung und Nutzung der Twitter-Daten wurde im Rahmen eines Datasprints im Anschluss an die Tagung »Nachhaltige Archivierung, Erschließung, Bereitstellung dynamischer Daten aus sozialen Medien – Twitter und danach«[32] im März 2024 erstmals ermöglicht. Das Korpus ist außerdem Teil des Digital-Humanities-Calls 2024 der DNB.[33] Die Nutzung des Korpus ist grundsätzlich nur in den Räumlichkeiten und auf der Infrastruktur und den Geräten der DNB möglich. Die Rechner, an denen mit den Twitter-Daten gearbeitet wird, haben aus rechtlichen Gründen keine Verbindung zum Internet. Da für die Nutzung je nach Anwendungsfall und Forschungsfrage bestimmte Software-Tools notwendig sind, die aus den o. g. Gründen vorinstalliert werden müssen, ist ein direkter Zugang zu den Daten nicht möglich, vielmehr eine gewisse Vorbereitung notwendig. Wegen der rechtlichen Einschränkungen kann es zudem sein, dass bestimmte Ergebnisse der Arbeit mit den Daten nicht veröffentlicht werden dürfen, dies ist im Einzelfall zu prüfen. Interessierte müssen zudem im Umgang mit den Tools geschult sein, da die DNB keine Einführungen und Schulungen anbieten kann. Aus diesen Gründen ist für künftige Bereitstellungen und Nutzungen eine Bewerbung notwendig, die das konkrete Vorhaben und die Bedarfe der Interessierten beschreibt, sodass die DNB im Rahmen ihrer Kapazitäten im Dialog mit den Interessierten den Zugang ermöglichen kann. Viele Forschungsfragen lassen sich mit abgeleiteten Daten oder statistischen Kennzahlen zu aggregierten Daten bearbeiten, zum Beispiel die Häufigkeit bestimmter Hashtags über die Zeit, die Anzahl von Tweets für einen oder mehrere Hashtags, die häufigsten verlinkten Webseiten oder Eigennamen von Personen oder Städten.[34] Um unter diesen einschränkenden Bedingungen für den Zugang zum Gesamtkorpus dennoch einen direkten Zugang zu bestimmten Aspekten der Twitter-Sammlung zu bieten, sollen perspektivisch abgeleitete Daten definiert, hergestellt und über das DNBLab[35] frei zugänglich bereitgestellt werden.

Anmerkungen

1 PASSIG, Kathrin. Den Heuhaufen archivieren. In : RICHTER, Sandra (Hg.), *#LiteraturArchivDerZukunft / herausgegeben von Sandra Richter*. Bd. 173/174. Marbach am Neckar : Deutsche Schillergesellschaft, 2021, o. P. PASSIG, Kathrin. *Rucksack oder Rechenzentrum* [online]. 2022. [Zugriff am: 19. Juli 2023]. Verfügbar unter: https://www.dhd2022.de/closing-keynote/ ; BRIL, Marijn. *Performatively Archiving the Early Web: One Terabyte of Kilobyte Age* [online]. Sound & Vision, September 2023, Bd. 12, Nr. 23, S. 69. DOI 10.18146/view.293

2 BRUNS, Axel und WELLER, Katrin. Twitter as a First Draft of the Present – and the Challenges of Preserving It for

the Future. In : NEJDL, Wolfgang, HALL, Wendy, PARIGI, Paolo, u. a. (Hg.), *WebSci '16 : Proceedings of the 8th ACM Conference on Web Science* [online]. New York : Association for Computing Machinery (ACM), 2016, S. 183–189. DOI 10.1145/2908131.2908174 ; VLASSENROOT, Eveline, CHAMBERS, Sally, LIEBER, Sven, u. a. Web-Archiving and Social Media: An Exploratory Analysis: Call for Papers Digital Humanities and Web Archives – A Special Issue of International Journal of Digital Humanities. *International Journal of Digital Humanities* [online]. November 2021, Bd. 2, Nr. 1–3, S. 107–128. DOI 10.1007/s42803-021-00036-1 ; TSOLAK, Dorian, KNAUFF, Stefan, KÜHNE, Simon, u. a. *X-GTA: The Cross-Topic German Twitter Archive* [online]. preprint. [S. l.]: SocArXiv, 16 Juni 2023. [Zugriff am: 3. Juli 2023]. DOI 10.31235/osf.io/9tbd4

3 Mit umfassendem Konzept, aber sammlungsorientiert thematisch eingeschränkt etwa am Archiv der Sozialen Demokratie, WALZ, Annabel und MARQUET, Andreas (Hg.), *Sicher sichern? Social Media-Archivierung aus rechtlicher Perspektive im Archiv der sozialen Demokratie*. Archiv der Sozialen Demokratie 2022; mit umfangreicher multilingualer Sammlung und im Überschneidungsbereich von forschungsbasierter und institutioneller Archivierung am Forschungsinstitut GESIS siehe FAFALIOS, Pavlos, IOSIFIDIS, Vasileios, NTOUTSI, Eirini, u. a., *TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets*. Bd. 10843 [online]. 2018, S. 177–190. DOI 10.1007/978-3-319-93417-4_12

4 So beispielsweise SCHEFFLER, Tatjana, A German Twitter Snapshot. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* [online]. Reykjavik, Iceland: European Language Resources Association (ELRA), Mai 2014, S. 2284–2289. [Zugriff am: 24. April 2023]. Verfügbar unter: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1146_Paper.pdf; außerdem TSOLAK u. a., *X-GTA*.

5 KUPFERSCHMIDT, Kai. Twitter's Plan to Cut off Free Data Access Evokes »Fair Amount of Panic« among Scientists. *Science* [online]. Februar 2023. DOI 10.1126/science.adh0813 ; CALMA, Justine. Scientists Say They Can't Rely on Twitter Anymore. *The Verge* [online]. Mai 2023. [Zugriff am: 1. Juni 2023]. Verfügbar unter: https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research ; STOKEL-WALKER, Chris. Twitter Is Making Researchers Delete Data It Gave Them Unless They Pay $42,000. *inews.co.uk* [online]. Mai 2023. [Zugriff am: 25. Mai 2023]. Verfügbar unter: https://inews.co.uk/news/twitter-researchers-delete-data-unless-pay-2364535 ; TAYLOR, Bryn. *Letter: Twitter's New API Plans Will Devastate Public Interest Research* [online]. Coalition for Independent Technology Research, 3. April 2023. [Zugriff am: 21. Mai 2024]. Verfügbar unter: https://independenttechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/

6 SCHWEIGER, Wolfgang. Musk führt Lese-Limit für Tweets ein. *Deutschlandfunk* [online]. Juli 2023. [Zugriff am: 21. Mai 2024]. Verfügbar unter: https://www.deutschlandfunk.de/elon-musk-twitter-limit-tweets-lesebeschraenkung-einschraenkung-100.html

7 KLIMPEL, Paul und KEIPER, Jürgen. *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*. Berlin 2013. [Zugriff am: 20. Mai 2024]. Verfügbar unter: https://irights-media.de/publikationen/was-bleibt-nachhaltigkeit-der-kultur-in-der-digitalen-welt/ ; PASSIG, Kathrin. *Rucksack oder Rechenzentrum* [online]. 2022. [Zugriff am: 19. Juli 2023]. Verfügbar unter: https://www.dhd2022.de/closing-keynote/

8 PASSIG, Heuhaufen.

9 WALZ und MARQUET, *Sicher sichern?*; SCHLESINGER, Claus-Michael und Mona ULRICH. Quelltexte in Netzliteratur aus archivarischer und literaturwissenschaftlicher Perspektive. In: Madeleine BROOK, Stefanie HUNDEHEGE und Caroline JESSEN (Hrsg.), *»Verschwinden« Vom Umgang mit materialen & medialen Verlusten in Archiv und Bibliothek*. Göttingen: Wallstein. 2024. DOI 10.15499/KDS-004-009 ; SCHLESINGER und ULRICH, Quelltexte.

10 LIBRARY OF CONGRESS. *Update on the Twitter Archive at the Library of Congress* [online]. 2017. [Zugriff am: 29. April 2024]. Verfügbar unter: https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf ; VLASSENROOT u. a., Web-Archiving and Social Media ; PEHLIVAN, Zeynep, THIÈVRE, Jérôme und DRUGEON, Thomas. Archiving Social Media: The Case of Twitter. In: GOMES, Daniel, DEMIDOVA, Elena, WINTERS, Jane, u. a. (Hg.), *The Past Web: Exploring Web Archives* [online]. Cham: Springer International Publishing, 2021, S. 43–56; DOI 10.1007/978-3-030-63291-5_5

11 LIBRARY OF CONGRESS, *Update*; OSTERBERG, Gayle. *Update on the Twitter Archive at the Library of Congress | Timeless* [online]. The Library of Congress, 26 Dezember 2017. [Zugriff am: 29. April 2024]. Verfügbar unter: https://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2

12 VLASSENROOT u. a., Web Archiving and Social Media.

13 So listet etwa der kollaborativ gepflegte Katalog der Gruppe Documenting the Now 143 Datensätze zur weiteren Verwendung. DOCNOW COLLABORATORS, *DocNow Tweet Catalog* [online]. [o. J.]. [Zugriff am: 21. Mai 2024]. Verfügbar unter: https://catalog.docnow.io/

14 SCHEFFLER, Tatjana. A German Twitter Snapshot. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* [online]. Reykjavik, Iceland: European Language Resources Association (ELRA), Mai 2014, S. 2284–2289. [Zugriff am: 24. April 2023]. Verfügbar unter: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1146_Paper.pdf ; TSOLAK u. a., *X-GTA*.

15 FAFALIOS, Pavlos, IOSIFIDIS, Vasileios, NTOUTSI, Eirini, u. a. *TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets*. Bd. 10843 [online]. 2018, S. 177–190. [Zugriff am: 25. April 2023]. DOI 10.1007/978-3-319-93417-4_12 ; PFEFFER, Juergen, MATTER, Daniel, JAIDKA, Kokil, u. a. *Just Another Day on Twitter: A Complete 24 Hours of Twitter Data* [online]. 26 Januar 2023. [Zugriff am: 1. März 2023]. Verfügbar unter: http://arxiv.org/abs/2301.11429

16 HAMMER, Luca. *Vermessung der deutschsprachigen Twittersphäre*. Universität Paderborn, 28. Dezember 2020. Verfügbar unter: https://lucahammer.com/ba2020

17 PASSIG, Rucksack oder Rechenzentrum; mit ähnlicher Forderung aus Forschungsperspektive BRUNS und WELLER, Twitter as a First Draft of the Present.

18 PASSIG, Heuhaufen.

19 BRUNS und WELLER, Twitter as a First Draft of the Present.

20 ESPENSCHIED, Dragan. Digital Objecthood. In: ARTUT, Selcuk, KARAMAN, Osman Serhat und YILMAZ, Cemal (Hg.), *TECHNOLOGICAL ARTS PRESERVATION*. Istanbul, 18 Juni 2021, S. 116–139; ESPENSCHIED, Dragan und RECHERT, Klaus. *207.1 Fencing Apparently Infinite Objects.* [online]. Open Science Framework, August 2022. DOI 10.17605/OSF.IO/6F2NM

21 HAMMER, Vermessung.

22 SCHEFFLER, A German Twitter Snapshot.

23 TSOLAK u. a., X-GTA.

24 Wir danken Luca Hammer für die maßgebliche Unterstützung bei der Entwicklung des Suchstrings.

25 TWITTER. *Evaluating Language Identification Performance* [online]. 16 November 2015. [Zugriff am: 11. März 2022]. Verfügbar unter: https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance

26 Siehe aber für Einschätzungen und methodischen Umgang mit diesem Problem SCHEFFLER, A German Twitter Snapshot und TSOLAK u. a., X-GTA; für das Niederländische außerdem mit umfangreichen Vergleichsverfahren KREUTZ, Tim und DAELEMANS, Walter, How to Optimize Your Twitter Collection: Dutch Keywords for Better Coverage. *Computational Linguistics in the Netherlands Journal* [online]. Dezember 2019, Bd. 9, S. 55–66. Verfügbar unter: https://clinjournal.org/clinj/article/view/92

27 PFEFFER, Juergen, MOOSEDER, Angelina, LASSER, Jana, u. a. *This Sample Seems to Be Good Enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API* [online]. 11 April 2023. [Zugriff am: 24. April 2023]. DOI 10.48550/arXiv.2204.02290

28 SCHLESINGER, Claus-Michael, WOLDERING, Britta. *Tweets archivieren* [online]. 2023. DOI 10.5281/zenodo.8006204

29 DOCUMENTING THE NOW. *Twarc* [online]. Twarc Documentation, [o. J.]. [Zugriff am: 19. Juli 2023]. Verfügbar unter: https://twarc-project.readthedocs.io/en/latest/

30 Für die mit auf Archivierung optimierten Twarc-Einstellungen und entsprechende Metadaten siehe die Twarc-Dokumentation DOCUMENTING THE NOW, *Twarc*. Eine Dokumentation der einzelnen Datenpunkte wird in Zukunft für die Bereitstellung verfügbar gemacht und in den Forschungsdaten nachgetragen. SCHLESINGER, Claus-Michael, WOLDERING, Britta, *Forschungsdaten Tweet-Archivierung DNB Und SDC-4Lit 2023* [online]. 2024. DOI 10.5281/zenodo.11278342

31 SCHEFFLER 2014, vgl. TSOLAK 2023.

32 Tagungsdokumentation und Präsentationen: https://wiki.dnb.de/x/AAgYF

33 Digital Humanities Calls der DNB: https://www.dnb.de/DE/Professionell/Services/WissenschaftundForschung/DHCall/dhcall_node.html

34 Für eine beispielhafte Umsetzung dieses Ansatzes mit einem Tweet-Korpus siehe FAFALIOS u. a., Tweetskb.

35 https://www.dnb.de/dnblab

Verfasser*innen

Dr. Claus-Michael Schlesinger,
Kompetenzwerkstatt Digital Humanities,
DFG-Projekt Furesh II, Humboldt-Universität zu Berlin, Universitätsbibliothek,
Unter den Linden 6, 10099 Berlin,
Telefon +49 30 2093-99642,
claus-michael.schlesinger@hu-berlin.de

Dr. Britta Woldering, Automatische Erschließungsverfahren, Netzpublikationen,
Deutsche Nationalbibliothek, Adickesallee 1,
60322 Frankfurt am Main,
Telefon +49 69 1525-1541,
b.woldering@dnb.de
Foto: Sarah Kastner