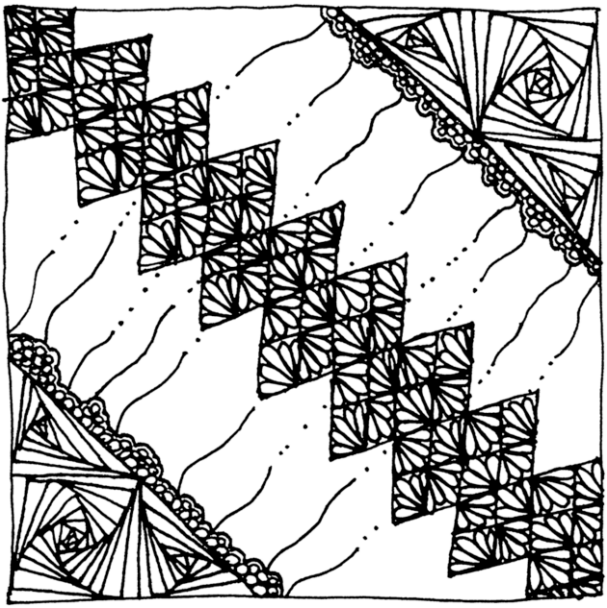
Datenquellen

2



J. Jünger, C. Gärtner, *Computational Methods für die Sozial- und Geisteswissenschaften*, https://doi.org/10.1007/978-3-658-37747-2_2

Zusammenfassung

Dieses Kapitel beinhaltet eine Einführung in Datenquellen, aus denen Daten mit Computational Methods gewonnen und analysiert werden können. Sie lernen, worin sich verschiedenen Verfahren automatisierter Datenerhebung unterscheiden und wie Ressourcen im Web identifiziert werden. Außerdem erfahren Sie, wo wissenschaftlich verwertbare Daten zu finden sind.

Im Online-Repositorium unter https://github.com/strohne/cm finden Sie begleitend zum Kapitel weitere Materialien, auf die wir im Text mit ☛ verweisen.

Schlüsselwörter

Uniform Resource Locator (URL) · Application Programming Interface (API) · Repositorien · Datenspenden · Open Data

Im Bereich Computational Methods besteht eine zentrale Frage darin, wie automatisiert sozial- und geisteswissenschaftlich relevante Daten gewonnen werden können. Automatisierte Datenerhebungen sind vor allem dort sinnvoll, wo sehr viele Daten anfallen, die man ungern manuell aufbereiten will. Das trifft beispielsweise für die Kommunikation auf Online-Plattformen wie Facebook zu. Die Daten sind dabei in der Regel schon vorhanden und werden nicht erst von Wissenschaftler:innen erstellt. Sie fallen im Zusammenhang mit menschlichem Verhalten ohnehin an, deshalb spricht man auch von prozessgenerierten Daten (Johnson und Turner 2003). Dennoch kann man nicht davon ausgehen, dass hier ein unverfälschter Zugang zu einer ohnehin vorhandenen sozialen Wirklichkeit besteht, vielmehr wird Wirklichkeit erst auf den Plattformen erzeugt – zum Beispiel spielen die Funktionen der Plattformen eine wichtige Rolle dafür, welche Daten entstehen und zugänglich sind (siehe zum Beispiel Jünger 2021).

Über verschiedene Datenzugänge werden unterschiedliche Repräsentationen von Wirklichkeit sichtbar. Dieser Punkt lässt sich gut am Beispiel von Online-Kommunikation verdeutlichen. Aus technischer Sicht ist Online-Kommunikation dadurch gekennzeichnet, dass zwei Maschinen miteinander interagieren. Auf der Seite der Nutzer:innen wird diese Maschine als Client bezeichnet, der Client schickt eine Anfrage an einen Server. Der Server bearbeitet die Anfrage und schickt eine Antwort zurück. Beim Surfen im Web findet dieses Wechselspiel ausgehend von einem Browser wie Firefox, Chrome oder Safari statt. Automatisierte Datenerhebung ist nun dadurch gekennzeichnet, dass Skripte oder Programme eingesetzt

werden, um Daten beim Server anzufragen. Statt also die Adresse https://www.google.de in den Browser einzugeben, wird diese Adresse in einem Erhebungstool erzeugt und das Ergebnis wird weiterverarbeitet (Abb. 2.1).

Normalerweise antworten Webserver mit HTML-Dateien, die dann im Browser grafisch dargestellt werden. Diese HTML-Dateien enthalten die Daten und Verweise auf weitere Dateien, die zur Darstellung benötigt werden, etwa zu Bilddateien. Formatvorlagen in der Form von CSS-Dateien steuern darüber hinaus die Gestaltung, etwa die Schriftfarbe, und JavaScript-Dateien fügen interaktive Elemente hinzu, zum Beispiel zum Auf- und Zuklappen von Menüs. Wenn bei der automatisierten Datenerhebung mit diesen HTML-Dateien gearbeitet wird, spricht man von Webscraping (siehe Abschn. 7.1). Die Dateien werden hierbei nicht angezeigt, sondern es werden einzelne Daten wie Texte oder Tabellen aus dem HTML-Quelltext von Webseiten extrahiert.

Viele Webseiten, unter anderem Social-Media-Plattformen, stellen zusätzlich sogenannte Application Programming Interfaces (APIs) zur Verfügung. Eine solche API unterscheidet sich von Webseiten dadurch, dass sie für den automatisierten Zugriff unabhängig von einem Browser entwickelt wird. Während Anbieter die Struktur einer Webseite bei Bedarf unangekündigt ändern – vor allem, wenn neue Funktionen eingeführt werden –, garantieren die Betreiber von APIs in der Regel,

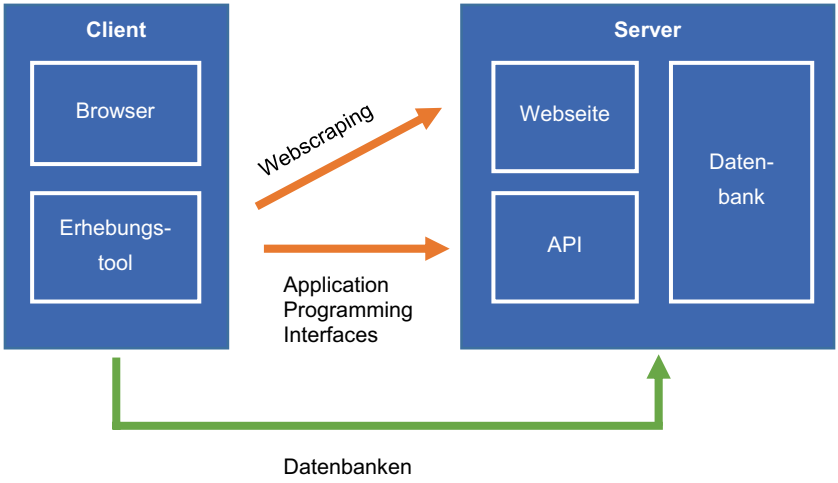


Abb. 2.1 Verfahren automatisierter Datenerhebung im Web. (Quelle: eigene Darstellung)

dass diese über lange Zeiträume stabil bleiben. Nur deshalb lohnt sich für Drittanbieter die Investition in eigene Apps, die auf Fremddaten aufbauen. Würde sich etwa die Struktur der Google-Maps-API immer wieder ändern, gäbe es keine Garantie, dass eine darauf aufbauende Geocaching-App danach noch funktioniert. Ein weiterer Unterschied besteht darin, dass die zurückgegebenen Daten nicht im HTML-Format, sondern in stärker vorstrukturierten und damit leichter verarbeitbaren Formaten wie JSON verschickt werden (siehe Kap. 3).

Sowohl der Zugang über Webscraping als auch die Nutzung von APIs sind davon abhängig, welche Daten ein Betreiber zugänglich macht. Im Endeffekt vermitteln beide Wege den kontrollierten Zugriff auf die Datenbanken des Anbieters, es handelt sich dabei jedoch um verschiedene Repräsentationen der Daten. In seltenen Fällen werden die Datenbanken selbst zur Verfügung gestellt. Ein Beispiel wäre die Wikipedia-Datenbank. Auch hier erfolgt der Zugriff aber nicht unvermittelt, zum einen muss die Datenbank erst heruntergeladen werden und zum anderen ist für die Arbeit mit Datenbanken ein passendes Datenbankmanagementsystem (DBMS) nötig (siehe Kap. 3). Insofern gibt es keine unvermittelten Daten und es muss stets reflektiert werden, wofür die erhobenen Daten stehen. In den folgenden Kapiteln werden Grundlagen zu den drei verschiedenen Datenzugängen und entsprechenden Datenformaten vermittelt sowie einige typische Datenquellen benannt.

2.1 Webseiten

Ein wesentliches Element des Web sind Uniform Resource Locators (URLs), mit denen die verschiedenen Ressourcen im Web adressiert werden. Eine solche URL besteht in der Regel aus fünf Komponenten (Abb. 2.2, ☛ *Repositorium*):

- Das **Protokoll** gibt an, in welcher Sprache die beiden Computer miteinander interagieren. Im Web wird dafür typischerweise HTTP (Hypertext Transfer Protokoll) oder für verschlüsselte Kommunikation HTTPS verwendet.
- Die **Domain** identifiziert den Server und besteht selbst wieder aus mehreren durch Punkt getrennten Teilen. Im Beispiel ist die Top-Level-Domain „com“, die Domain selbst ist „youtube“ und das Präfix „www“ wird als Subdomain bezeichnet.
- Der **Pfad** gibt die Operation oder die Ressource auf dem Server an. Vereinfachend kann man davon ausgehen, dass hinter einer Domain ein Computer steht und der Pfad die Datei oder das Programm angibt, auf welche zugegriffen werden soll.

- Nach einem Fragezeichen folgen die **Parameter der Anfrage** (engl. *query string*), um dem Server genauere Angaben zum Auffinden der Ressource mitzugeben. Ein Parameter besteht immer aus einem Namen, auf den nach einem Gleichheitszeichen ein Wert folgt. Mehrere Parameter werden durch das &-Zeichen getrennt.
- Das **Hashfragment** am Ende der URL wird niemals an den Server gesendet, sondern lediglich im Browser ausgewertet. Damit werden einzelne Teile der Webseite angesprochen, zum Beispiel der Kommentarbereich, sodass die Anzeige direkt zu diesem Abschnitt springen kann.

Darüber hinaus können in Internetadressen noch weitere Angaben vorkommen, die für die Bearbeitung der Anfrage relevant sind, beispielsweise ein sogenannter Port oder Zugangsdaten.

Zu beachten ist, dass einige Zeichen in URLs speziell kodiert werden müssen – das betrifft die für den Aufbau der URL reservierten Zeichen wie das Fragezeichen, aber auch Umlaute. Diese Zeichen werden durch Prozentkodierung angegeben, sodass aus dem Umlaut ä beispielsweise %C3%A4 wird. Für das Leerzeichen gibt es zwei Varianten, zum einen kann universell die Prozentkodierung %20 verwendet werden, zum anderen ist innerhalb der Parameter das Pluszeichen + anzutreffen.

Welche Rolle spielen URLs nun für Daten auf Webseiten? Im einfachsten Fall ist eine Ressource durch eine URL abrufbar und die entsprechende Webseite enthält Daten wie zum Beispiel eine eingebettete Tabelle mit Mitgliederzahlen politischer Parteien. Auch die gesamte Webseite kann Gegenstand der Analyse sein, wenn Blogs oder Nachrichtenseiten untersucht werden sollen. Zudem sind die URLs selbst für wissenschaftliche Analysen von Interesse, denn dadurch können die Verbindungen zu anderen Seiten verfolgt werden, um so Netzwerke zwischen Webseiten und Akteuren nachzuvollziehen.

Abb. 2.2 Die Bestandteile einer URL. (Quelle: eigene Darstellung)

Im Web findet sich eine Vielzahl an Webseiten, auf denen Daten beispielsweise in Tabellen- oder Listenform bereitgestellt werden:

- Primärdaten finden sich insbesondere in Registern politisch-administrativer Angebote. Eine Anlaufstelle für Listen von Unternehmen oder Vereinen ist in Deutschland das gemeinsame Registerportal der Länder (Ministerium der Justiz Nordrhein-Westfalen 2020). Auch das statistische Bundesamt stellt Daten zur Verfügung (Destatis 2022).
- Berufsverbände und andere Interessenvertretungen listen häufig Daten über ihre Mitglieder auf. So lassen sich Medienangebote unter anderem über die Informationsgesellschaft zur Feststellung der Verbreitung von Werbeträgern (IVW 2022) oder über den Bundesverband Digitalpublisher und Zeitungsverleger (BDZV 2022) identifizieren.
- Themenportale und Plattformen bieten häufig Überblicksseiten über ihre Datenbestände an. Das reicht von Medienangeboten wie Fernsehserien oder Podcasts über Fußballergebnisse bis zu Cocktailrezepten und App-Stores.
- Aufbereitete Daten zu allen möglichen Themen finden sich auch in den Artikeln von Online-Enzyklopädien und Nachschlagewerken wie Wikipedia (Wikimedia Deutschland 2022) oder Fandom (Fandom 2022).

Beim Zugang zu diesen Daten stößt man auf ganz unterschiedliche Rahmenbedingungen. Im einfachsten Fall sind alle Daten auf einer einzelnen über eine URL identifizierbare Seite enthalten, zum Beispiel in einer Wikipedia-Tabelle (siehe Abschn. 7.1). Häufig werden lange Listen aber auch über mehrere Seiten verteilt, wobei jede Seite eine eigene URL erhält. Beobachten Sie beim Surfen im Web die URLs: Typischerweise wird die Seite über einen Parameter wie `page=5` angegeben und Protokoll, Domain sowie Pfad bleiben gleich.

Die Paginierung, das heißt die Aufteilung auf mehrere Seiten, wird auf stark interaktiven Seiten jedoch nicht immer in der Adressleiste sichtbar, sondern die einzelnen Seiten werden beim Scrollen nach und nach über JavaScript und sogenannte XMLHttpRequests nachgeladen, was die automatisierte Erhebung erschwert. Eine weitere Hürde sind Datenbanken, in denen über Suchformulare recherchiert wird. Hier reicht die Angabe einer URL nicht aus, sondern die Suchbegriffe müssen auf anderem Weg an den Server übermittelt werden. Das HTTP-Protokoll sieht verschiedene Methoden der Interaktion mit einem Webserver vor: GET-Anfragen rufen eine URL auf, POST-Anfragen senden weitere Nutzdaten an diese Adresse und DELETE-Anfragen sind zum Löschen von Daten vorgesehen. Suchfunktionen bauen häufig auf POST-Anfragen auf. Da die Daten nicht über Links identifizierbar sind, können solche Inhalte auch nicht einfach über

Suchmaschinen wie Google gefunden werden, wofür sich der Begriff Deep Web eingebürgert hat.

Welche Anfragen genau beim Surfen an einen Server geschickt werden, lässt sich gut mit der Entwicklerkonsole des Browsers nachvollziehen, die in den meisten Browsern mit der Taste F12 aktiviert wird (siehe das Beispiel zum Extrahieren von URLs in Abschn. 4.1). Der praktische Umgang mit den verschiedenen Formen von Webscraping wird in Abschn. 7.1 vermittelt. An dieser Stelle sind zunächst drei Hinweise auf ethisch-rechtliche Voraussetzungen wichtig. Erstens enthalten die Nutzungsbedingungen (engl. *terms of services*) von Webseiten und insbesondere von Social-Media-Plattformen Regelungen, mit denen sich die Betreiber eine automatisierte Datenerhebung häufig verbitten. Allerdings gehören insbesondere Suchmaschinen wie Google, die solche Erhebungen systematisch durchführen, selbstverständlich zum Web dazu, diese erfassen über automatisiertes Crawling die Inhalte von anderen Webseiten. Einen Einblick, welche Regelungen eine Webseite dafür vorsieht, können Sie sich über die robots.txt verschaffen. Diese Datei ist in der Regel auf jedem Server verfügbar und kann abgerufen werden, indem der Name an den Domainnamen angehängt wird, probieren Sie es zum Beispiel bei Facebook aus: https://www.facebook.com/robots.txt. Zweitens finden sich in der Datenschutzgrundverordnung und im Urheberrecht spezielle Regelungen für wissenschaftliche Zwecke, die zum Beispiel Textmining unter bestimmten Umständen explizit erlauben. Drittens sind ethische Abwägungen notwendig, insbesondere, wenn personenbezogene Daten betroffen sind. Das bedeutet: Die Möglichkeiten und Grenzen automatisierter Erhebungen müssen für jedes Projekt im Einzelfall reflektiert werden. Orientierung geben Ethikkodizes und Handreichungen (zum Beispiel RatSWD 2019) und der rege Diskurs in der Forschungsliteratur (unter anderem Bruns 2019; Fiesler et al. 2020; Kotsios et al. 2019; Thelwall und Stuart 2006).

2.2 Application Programming Interfaces

Das Extrahieren von Daten auf Webseiten baut zwar auf Standards im Web auf, erfordert jedoch mitunter eine detaillierte Auseinandersetzung mit der Struktur der Webseiten. Zudem ist Webscraping besonders von einigen Social-Media-Plattformen wie Facebook, Twitter oder YouTube laut deren Nutzungsbedingungen nicht erwünscht. Webseitenbetreiber bauen mitunter Hürden ein, um das Webscraping zu erschweren. Ein bereits vorstrukturierter Datenzugang ist aber bei vielen Plattformen mittels Application Programming Interfaces (APIs) möglich, über die Webseitenbetreiber kontrollieren, wer wie viele und welche Daten erheben kann.

Ganz allgemein legen Programmierschnittstellen fest, wie zwei Programme miteinander interagieren können (Jacobson et al. 2012, S. 5). Diese Schnittstellen sind meistens nicht vorrangig für wissenschaftliche Datenanalysen eingerichtet worden, sondern für die Entwicklung von Drittanwendungen. Im Web wird auf diese Weise beispielsweise die Funktion umgesetzt, dass man sich auf anderen Seiten „Mit Google einloggen“ kann. Die API-Anbieter legen dazu Endpunkte und Parameter fest, auf die andere Programme, sogenannte API-Konsumenten, zugreifen. Ein Endpunkt ist einfach eine URL wie `https://api.twitter.com/2/users/`. Diese URL wird um weitere Pfad- und Queryparameter ergänzt, mit denen etwa die öffentlichen Profilinformationen einzelner Nutzer:innen abgefragt werden. Pfadparameter wie `show.json` werden direkt an den Pfad angehängt, wohingegen Queryparameter wie `?screen_name=wissen_lockt` als Liste von Name-Wert-Paaren nach einem Fragezeichen angegeben werden. Im Gegensatz zu einer normalen Webseite geben APIs in der Regel nicht HTML, sondern deutlich leichter verarbeitbare JSON-Formate zurück (Abb. 2.3 und Kap. 3).

Eine API ist mehr als eine Software, sie ist ein Vertrag zwischen dem API-Anbieter und dem API-Konsumenten. Der Anbieter sichert damit zu, dass der Zugriff über einen längeren Zeitraum bestehen bleibt und der Konsument sorgt letztendlich dafür, dass sich die Dienste des Anbieters in der Welt verbreiten. So wie sich Hersteller von USB-Sticks darauf verlassen, dass die USB-Buchse immer die gleichen Abmessungen haben, verlassen sich Anwendungsentwickler darauf, dass sich die Endpunkte und die Datenformate nicht ändern. Wichtig ist deshalb die genaue Dokumentation (engl. *reference)* der API, in der Endpunkte, Parameter und Rückgabeformate beschrieben werden. Die Betreiber stellen die Dokumentation mehr oder weniger übersichtlich zusammen, darüber hinaus kommen auch maschinenlesbare Standards wie OpenAPI[1] zum Einsatz.

Viele für wissenschaftliche Analysen verwendete APIs bauen auf REST-Prinzipien (Fielding 2000) auf, das heißt, einzelne Ressourcen sind wie Webseiten über URLs ansprechbar. Einige APIs sind so weit standardisiert, dass die Endpunkte immer auf die gleiche Art und Weise aufgebaut sind oder dass auch die Dokumentation der API über die API selbst abgerufen werden kann. Beispielsweise folgen die zum Abgleich heterogener Datenbestände eingesetzten Reconci-

[1] Siehe OpenAPI (2022; https://github.com/OAI/OpenAPI-Specification).

URL der Webseite:

https://twitter.com/wissen_lockt

URL des API-Endpunkts:

https://api.twitter.com/1.1/users/show.json?screen_name=wissen_lockt

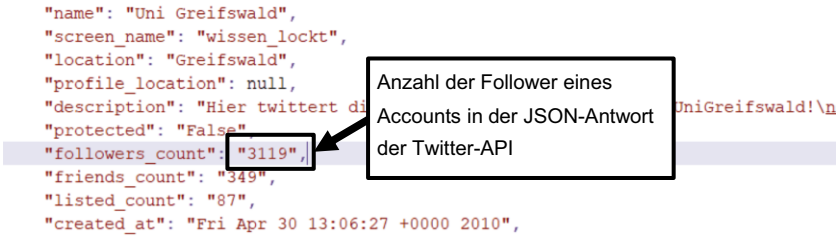


Abb. 2.3 Inhalt einer Webseite (HTML) und Antwort einer API (JSON) im Vergleich. (Quelle: eigene Darstellung)

liation Service APIs[2] einer übergeordneten Spezifikation und können so in Tools wie OpenRefine[3] verwendet werden. OpenRefine ermöglicht es, über die Einbindung mehrerer APIs etwa eine Liste von Personen oder Unternehmen gleichzeitig mit einem Register politisch sanktionierter Akteure und mit Einträgen bei der Deutschen Nationalbibliothek abzugleichen.[4] Auch die im Semantic Web einge-

[2] Siehe Entity Reconciliation Community Group (2022; https://reconciliation-api.github.io/specs/0.1/).

[3] Siehe Huynh (2022; https://openrefine.org/).

[4] Für eine Liste von APIs, die in OpenRefine eingebunden werden können, siehe OpenRefine (2021; https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources).

setzten APIs folgen Standards wie Hydra[5] oder stellen nach einem festgelegten Schema sogenannte SPARQL-Endpunkte bereit (siehe Kap. 3). Trotz dieser Standardisierungen bleibt eine Auseinandersetzung mit den speziellen Endpunkten einer API nicht aus.

Wie auch auf Webseiten setzen viele Anbieter eine Registrierung oder sogar eine Vorabprüfung des geplanten Projekts voraus. Insbesondere bei Social-Media-Plattformen wie Facebook oder YouTube ist der Zugang stark kontrolliert. Dagegen setzen sich Organisationen wie die Open Knowledge Foundation dafür ein, vor allem Daten öffentlich-rechtlicher Einrichtungen offen zugänglich zu machen, zum Beispiel über das Portal OffeneRegister.de. Diese Bestrebungen werden unter dem Schlagwort Open Data auch politisch aufgegriffen, im Jahr 2017 wurde vom Bundestag dazu das sogenannte Open-Data-Gesetz beschlossen (EGovG §12).

Ein Verzeichnis von APIs und weitere Erläuterungen dazu, was eine API ist, finden Sie auf ProgrammableWeb.[6] Darüber hinaus lohnt es sich stets zu prüfen, ob eine Webseite oder Plattform eine API anbietet, auch wenn dies nicht auf den ersten Blick erkennbar ist. Die Übergänge zwischen Webseiten und APIs sind fließend, weil Webanwendungen in vielen Fällen auf Ebene des dahinter liegenden Content Management Systems auf APIs aufbauen. So reicht es mitunter aus, einen Parameter in der URL einer Webseite zu ändern, um das Format von HTML auf JSON umzustellen (siehe die Übungsaufgabe am Ende des Kapitels).

APIs können für vielfältige geistes- und sozialwissenschaftliche Zwecke eingesetzt werden. Erstens stellen Social-Media-Dienste APIs bereit, mit denen die auf der Plattform erzeugten Inhalte (Posts, Kommentare, …) erhoben werden können. Daneben finden sich zweitens Dienste, die andere Daten sammeln und aggregieren, in Zitationsdatenbanken werden etwa die Literaturverweise von wissenschaftlichen Aufsätzen gesammelt. Drittens stellen Cloud-Computing-Anbieter wie Amazon, IBM, Google oder Microsoft über APIs Analysemöglichkeiten zum Beispiel zur automatisierten Bilderkennung bereit. Einen Überblick über einige APIs finden Sie in Tab. 2.1. Wie Sie selbst mit APIs arbeiten können, wird in Abschn. 7.2 erläutert.

[5] Siehe Hydra W3C Community Group (2018; http://www.hydra-cg.com/drafts/use-cases/2.api-documentation.md).

[6] Siehe Berlind (2022; https://www.programmableweb.com/).

Tab. 2.1 Beispiele für Anbieter von Application Programming Interfaces (APIs)

Anbieter	Datenbestände bzw. mögliche Einsatzbereiche
Plattformdaten	
Facebook API	Posts und Kommentare auf Facebook (Meta 2022; https://developers.facebook.com)
MediaWiki Action API	Daten der Wikimedia-Projekte, z. B. Wikipedia (MediaWiki 2022; https://www.mediawiki.org/wiki/API)
Twitter API	Tweets und Informationen zu Nutzer:innen (Twitter 2022d; https://developer.twitter.com/)
YouTube Data API	Videoinformationen und Kommentare auf YouTube (Google Developers 2022; https://developers.google.com/youtube/)
Aggregation von Daten	
Abgeordnetenwatch	Daten zu Wahlen und Abgeordneten (abgeordnetenwatch.de 2022; https://www.abgeordnetenwatch.de/api)
DWDS	Worthäufigkeiten und Wörterbucheinträge des Digitalen Wörterbuchs der deutschen Sprache (Berlin-Brandenburgische Akademie der Wissenschaften 2022; https://www.dwds.de/d/api)
GovData	Amtliche Daten, beispielsweise zu den Kategorien Bevölkerung, Gesundheit, Verkehr, Justiz (GovData 2022; https://www.govdata.de/)
Media Cloud	Online-Berichterstattung (Media Cloud 2022; https://mediacloud.org/)
OffeneRegister.de	Unternehmensdaten aus dem Handelsregister (Datasette 2022; https://db.offeneregister.de/openregister)
Open Citations	Zitationsdatenbank mit bibliografischen Informationen (Peroni und Daquino 2020; https://opencitations.net/index/api/v1)
Open Corporates	Weltweit gesammelte Unternehmensdaten (OpenCorporates 2022; https://api.opencorporates.com/)
Datenaufbereitung oder -analyse	
Amazon Web Services: Rekognition	Bild- und Videoanalyse (Amazon Web Services 2022a; https://aws.amazon.com/de/rekognition/)
Google Cloud Vision API	Erkennen von Objekten auf Bildern (Google 2022d; https://cloud.google.com/vision)
IBM Watson Machine Learning	Bild- und Textklassifikation (IBM 2022; https://www.ibm.com/de-de/cloud/machine-learning)
Microsoft Cognitive Services	Umwandlung von Audio- in Textdateien und Bilderkennung (Microsoft 2022a; https://docs.microsoft.com/en-us/azure/cognitive-services)
Perspective API	Analyse von Texten auf Hate-Speech (Perspective 2021; https://www.perspectiveapi.com/)

Quelle: Eigene Darstellung

2.3 Datenbanken und Datensätze

Sowohl beim Besuchen von Webseiten als auch beim Einsatz von webbasierten APIs wird der Zugang zu Datenbanken über Schnittstellen vermittelt, die auf dem HTTP-Protokoll aufbauen. Jede Anfrage gibt dabei einen kleinen Ausschnitt der Datenbank zurück. Für wissenschaftliche Studien finden sich im Web auch vollständige Datenbanken. Die Vollständigkeit hat aber ihren Preis: Die Dateien können sehr groß werden und sind nach der internen Logik des Anbieters strukturiert. Ein eindrucksvolles Beispiel ist die Global Database of Events Language and Tone.[7] In diesem Projekt werden im Viertelstundentakt weltweit Nachrichtenseiten mit automatischer Textanalyse ausgewertet, aggregiert und die Ergebnisse werden zum Download zur Verfügung gestellt. Für ein Jahr umfasst die Datenbank über zwei Terrabyte. Die Arbeit mit solch umfangreichen Datensätzen setzt spezifische Kenntnisse zum Umgang mit Datenbanken und auf mehrere Computer verteilte Systeme (Cloud Computing, siehe Abschn. 6.4) voraus. Doch nicht immer muss es sich um solche Datenmengen handeln, auch viele kleinere Datenbanken sind für sozial- und geisteswissenschaftliche Analysen hilfreich. Linguistische Korpora mit Chats oder WhatsApp-Nachrichten oder auch Listen mit den Einwohnerzahlen aller Länder der Welt sind vergleichsweise klein.

Zum Auffinden von Datensätzen eignen sich **Suchmaschinen** wie die Google Dataset Search oder Portale wie Kaggle (Tab. 2.2). Teilweise stellen Organisationen und Online-Plattformen ihre **Datenbanken** ganz oder in Teilen zur Verfügung, etwa die Wikipedia oder auch die International Movie Database. Auch Facebook macht ausgewählte Teile seiner Datenbanken für Wissenschaftler:innen verfügbar, zum Beispiel die meistgeteilten Links (URL dataset). Der Zugang ist in diesem Fall stark restringiert und erfolgt über die Organisationen Social Science One[8] oder Crowdtangle.[9] Eine besonders herausfordernde Datensorte stellen organisationsinterne **Verhaltensdaten** dar, sie umfassen beispielsweise Logdateien der Webseitennutzung, das Kaufverhalten in Online-Shops, die Bibliotheksnutzung oder die Daten von Fitness-Trackern. Der Zugang ist auf Mitarbeitende in den entsprechenden Organisationen bzw. Kooperationspartner beschränkt und nur mit starken datenschutzrechtlichen Schutzmaßnahmen möglich.

Zudem werden die Forschungsdaten wissenschaftlicher Studien zunehmend in öffentlichen **Repositorien** abgelegt, um eine Nachnutzung und Nachprüfung zu er-

[7] Siehe Leetaru (2021; https://gdeltproject.org/).

[8] Siehe Harvard University (2022b; https://socialscience.one/).

[9] Siehe Meta (2021; https://www.crowdtangle.com/).

Tab. 2.2 Beispiele für online verfügbare Datenbanken

Suchmaschinen, Repositorien und Verzeichnisse	
AMiner	Datensätze für soziale Netzwerkanalysen (CKCEST 2022; https://cn.aminer.org/data-sna)
CLARIN Virtual Language Observatory	Suchmaschine für Sprachdaten (CLARIN 2022; https://vlo.clarin.eu/)
GESIS	Repositorium für sozialwissenschaftliche Forschungsdaten (GESIS 2022; https://search.gesis.org)
Google Dataset Search	Suche nach öffentlichen Datensätzen (Google 2022f; https://datasetsearch.research.google.com)
Google Public Data	Verzeichnis mit Datensätzen (Google 2014; https://www.google.com/publicdata/directory)
Harvard Dataverse	Repositorium für Forschungsdaten (Harvard University 2022a; https://dataverse.harvard.edu/)
ICPSR datasets	Repositorium für Forschungsdaten (University of Michigan 2022; https://www.icpsr.umich.edu/)
Kaggle	Anlaufstelle für die Data-Science-Community, hostet eine Vielzahl an Datensätzen (Kaggle 2022; https://www.kaggle.com/)
Networkrepository	Repositorium mit wissenschaftlichen Netzwerkdaten (Rossi und Ahmed 2022; http://networkrepository.com/)
OSF	Repositorium für Forschungsdaten (Center for Open Science 2022; https://osf.io/)
Beispiele für kuratierte Datensätze und Korpora	
Blog Authorship Corpus	Blog-Posts auf blogger.com, sortiert nach Alter und Geschlecht der Autor:innen (Schler et al. 2005; https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus).
ConvoKit Datasets	Datensätze zur Analyse von Gesprächsverläufen und Dialogen (Chang et al. 2020; https://github.com/CornellNLP/Cornell-Conversational-Analysis-Toolkit)
COVID-19 Weibo Data	Zensierte und unzensierte chinesische Weibo-Nachrichten zu COVID-19 (Fu & Zhu 2020; https://doi.org/10.6084/m9.figshare.12199038)
DeReKo	Deutsches Referenzkorpus, weltweit größte Sammlung von Korpora mit belletristischen, wissenschaftlichen und journalistischen Texten der deutschen Gegenwartssprache (Kupietz et al. 2010; https://www.ids-mannheim.de/digspra/kl/projekte/korpora/).
DiDi-Korpus	Anonymisierte Kommentare und Chat-Nachrichten von Südtiroler Facebook-Nutzer:innen (Frey et al. 2019; https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/7)

(Fortsetzung)

Tab. 2.2 (Fortsetzung)

Beispiele für kuratierte Datensätze und Korpora	
Dortmunder Chat-Korpus	Chatprotokolle aus professionellen Kontexten und dem Freizeitbereich (Beißwenger 2013; Zugang über https://vlo.clarin.eu/)
D-Place	Database of Places, Language, Culture, and Environment. Ökonomische, soziodemographische und klimatische Bedingungen verschiedener Gesellschaftsformen zur Beschreibung der Menschheitsgeschichte (Jahre 1600–1990) (Kirby et al. 2016; https://d-place.org/)
FiveThirtyEight Russian Troll Tweets	Tweets von Accounts, die mit der Internet Research Agency (IRA) in Verbindung standen. Die russische IRA wird mit Einflussnahmen auf den amerikanischen Wahlkampf im Jahr 2016 in Verbindung gebracht (Linvill und Warren 2018; https://github.com/fivethirtyeight/russian-troll-tweets/)
GDELT	Laufend aktualisierte Datenbank mit weltweiter Berichterstattung. Die Texte durchlaufen eine Vielzahl automatisierter Inhaltsanalysen, zum Beispiel um Ereignisse zu extrahieren (Leetaru 2021; https://gdeltproject.org/)
IMDb	Datenbank zu Filmen und daran beteiligten Personen (Internet Movie Database 2022; https://www.imdb.com/interfaces/)
MoCoDa2	Mobile Communication Database, enthält WhatsApp-Konversationen, die von Nutzer:innen gespendet wurden (Beißwenger et al. 2020; https://db.mocoda2.de/)
Regesta Imperii	Herrscher- und Papsturkunden von den Karolingern bis hin zu Maximilian I. (751–1519) (Akademie der Wissenschaften und der Literatur Mainz 2022; http://www.regesta-imperii.de/)
Seshat	Historische Daten zur sozialen und politischen Organisation menschlicher Gesellschaften bis zur industriellen Revolution (Evolution Institute und Seshat Project 2021; http://seshatdatabank.info/)
Trump Tweets	Alle Tweets von Donald Trump (bis 20. Januar 2020) (Reese 2020; https://www.kaggle.com/austinreese/trump-tweets)
UCDP GED	Weltweite Beobachtung politischer Konflikte mit Angaben zu Todesfällen und beteiligten Akteuren (Uppsala Conflict Data Program 2016; https://ucdp.uu.se/)
useNews	Datensatz zur Verbreitung von Nachrichtenartikeln, enthält eine Kombination von Befragungsdaten, Metadaten von Artikeln und Daten zu Interaktionen auf Social-Media-Plattformen (Puschmann und Haim 2021; https://osf.io/uzca3/)

(Fortsetzung)

Tab. 2.2 (Fortsetzung)

Quellen für Trainingsmaterial	
GermEval SemEval	Veranstaltungen, bei denen im Rahmen von Aufgaben Datensätze zur automatischen Klassifikation bereitgestellt werden, etwa zu Offensive Language oder Sentiment-Analyse (GermEval 2019; https://germeval.github.io/; SemEval 2022; https://semeval.github.io/)
Oxford Text Archive	Literarische und linguistisch aufbereitete Texte, zum Beispiel zur Entwicklung automatischer Textanalyse (University of Oxford 2021; https://ota.bodleian.ox.ac.uk/repository/xmlui/)
TIGER Korpus	Textkorpus mit etwa 900.000 annotierten Token aus deutschen Nachrichtenartikeln (Brants et al. 2004; https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/).
Twitter Event Datasets	Sammlung von 30 unterschiedlichen Twitter-Datensätzen (2012–2016) zur Überprüfung der Replizierbarkeit von Twitter-Studien (Zubiaga 2018a; 2018b; https://figshare.com/articles/Twitter_event_datasets_2012-2016_/5100460)
UCI Machine Learning Repository	Trainingsdaten und Referenzdatensätze unter anderem für die Bild- und Texterkennung (Dua & Graff 2019; https://archive.ics.uci.edu/https://archive.ics.uci.edu/ml/datasets.php)
VoxCeleb	Ton- und Videoaufnahmen zum Beispiel für die Entwicklung von Sprach-, Gesichts- und Emotionserkennung (Visual Geometry Group 2022; http://www.robots.ox.ac.uk/~vgg/data/voxceleb/)
Wikipedia: List of datasets for machine-learning research	Auflistung von Trainingsdatensätzen für das Machine Learning, unter anderem zur Bilderkennung (Gesichts-Objekt-, Handschrifterkennung) und Texterkennung (Rezensionen, Nachrichten, Dialoge etc.) (Wikipedia 2022a; https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)

Quelle: Eigene Darstellung. Beachten Sie, dass die Datenqualität sehr unterschiedlich ausfällt, und vor der Verwendung der Daten eingeschätzt werden sollte

möglichen (Sekundärdatenanalyse). Eine zentrale Anlaufstelle für sozialwissenschaftliche Daten ist in Deutschland die GESIS. Mit dem Rückenwind der Open-Access-Bewegung fordern auch internationale Zeitschriften von ihren Autor:innen immer häufiger, dass die Daten verfügbar gemacht werden. Eine globale Plattform dafür betreibt die Open Science Foundation,[10] aber auch einzelne Universitäten bieten eigene Repositorien an. Zudem wird in Deutschland momentan mit viel Auf-

[10] Auch Sie selbst können Ihre Daten und Auswertungsskripte dort verfügbar machen. Wenn Sie selbst im Kontext von Qualifikationsarbeiten (B.A., M.A., Promotion) forschen, denken Sie darüber nach!

wand eine nationale Forschungsdateninfrastrukur aufgebaut,[11] über die Forschungsdaten nachhaltig nutzbar gemacht werden sollen. Die Anforderungen an solche wissenschaftlichen Datenbestände werden mit den FAIR-Prinzipien beschrieben – Findability, Accessibility, Interoperability und Reusability (Wilkinson et al. 2016).

Kuratierte Datensätze zu spezifischen Themen gehen teilweise auf wissenschaftliche Forschungsprojekte zurück oder werden von wissenschaftsnahen Organisationen für die Grundlagenforschung erstellt. Hierzu zählen beispielsweise von CLARIN bereit gestellte Sprachdaten, die von D-Place zusammengetragenen Strukturdaten zu menschlichen Gesellschaftsformen oder die vom UCDP erfassten Daten zu politischen Konflikten. Im Social-Media-Bereich sammelt beispielsweise Weiboscope Datensätze, mit denen sich die chinesische Zensur untersuchen lässt (Weiboscope 2022; zum Beispiel zu COVID-19, siehe Fu und Zhu 2020). Im Rahmen geisteswissenschaftlicher Grundlagenforschung werden insbesondere historische Daten in Langzeitprojekten erschlossen, die häufig an den Akademien der Wissenschaften angesiedelt sind. Diese Projekte verweisen auf eine lange Tradition, so hat die Aufarbeitung von königlichen und päpstlichen Urkunden des Mittelalters im Projekt Regesta Imperii bereits im Jahr 1829 begonnen (Akademie der Wissenschaften und der Literatur Mainz 2022). Regelmäßig werden in geisteswissenschaftlichen Projekten gedruckte Editionen veröffentlicht und die erschlossenen Daten werden zunehmend online in Datenbanken zur Verfügung gestellt.

Diese Datenbanken sind im Wesentlichen an zwei unterschiedlichen Zielstellungen ausgerichtet. Erstens erlauben einige Datenbestände inhaltliche Analysen, etwa um das Twitter-Verhalten von Donald Trump oder die Debatten im deutschen Bundestag auszuwerten. Zweitens stellen einige Projekte Daten als **Trainingsmaterial** für Machine-Learning-Verfahren bereit. Hier geht es darum, automatisierte Inhaltsanalysen zu entwickeln. Dazu gehören auch Korpora mit Videos für die Entwicklung automatischer Emotionserkennung oder Korpora mit Rezensionen für die Erkennung positiver und negativer Bewertungen. Ergänzend zu den inhaltlichen Daten sind diese Korpora manuell annotiert, das bedeutet zu jedem Video ist eine zusätzliche Angabe der Emotion oder zu jeder Rezension eine von Menschen vorgenommene Bewertung vorhanden. In jedem Fall sollten Sie sich genau mit der Qualität der Daten beschäftigen und darauf achten, dass deren Entstehung und Auswahl nachvollziehbar sind. So wie Sie keine Texte zitieren sollten, in denen die Aussagen nicht belegt oder begründet sind, sind auch undokumentierte Daten für wissenschaftliche Analysen ungeeignet.

Zum Erstellen annotierter Daten greifen einige Wissenschaftler:innen und Unternehmen auf **Crowdsourcing** zurück. Besonders bekannt, aber auch umstritten, ist zur Rekrutierung von Kodierer:innen für einfache Aufgaben die Plattform Ama-

[11] Siehe NFDI (2022; https://www.nfdi.de).

zon Mechanical Turk.[12] Annotierte Datensätze werden mitunter auf Veranstaltungen zum gemeinsamen Lernen (Hackathon oder Datathon genannt) oder bei Wettbewerben in Kooperation mit Unternehmen ausgegeben oder erstellt. Auf Plattformen wie Kaggle werden laufend Wettbewerbe zur Analyse von Datensätzen ausgeschrieben. Für die Analyse privater Kommunikation, etwa WhatsApp-Konversationen, sind Wissenschaftler:innen auf Datenspenden angewiesen (siehe zum Beispiel Beißwenger et al. 2020; Araujo et al. 2021).

Zusammenfassend unterscheiden sich die verschiedenen Datenzugänge also erstens danach, ob sie heterogene Datenbestände sammeln und durchsuchbar machen oder ob sie sich auf einzelne Themenbereiche beschränken. Zweitens werden Daten speziell für wissenschaftliche Zwecke erzeugt oder treten als Nebenprodukt von Handlungen auf. Die Verbreitung von informationstechnischen Systemen bringt es mit sich, dass umfangreiche Daten über menschliches Verhalten anfallen, mit denen alte, aber auch ganz neue Fragestellungen bearbeitet werden können. Drittens sind einige Datensätze nicht in erster Linie für inhaltliche Fragestellungen ausgelegt, sondern als Trainingsmaterial für die Methodenentwicklung. Viertens werden Datensätze nicht nur von wissenschaftlichen Einrichtungen mit entsprechenden Qualitätssicherungsverfahren, sondern auch von kommerziellen Anbietern bereit gestellt. Da letztere an marktwirtschaftlichen Prinzipien orientiert sind, kann es zu Interessenskonflikten kommen – die Datenqualität sollte vor der Verwendung besonders gründlich eingeschätzt werden. In Tab. 2.2 finden Sie einige Beispiele für die verschiedenen Arten von Datenquellen – verschaffen Sie sich selbst einen Eindruck davon, wie diese Daten einzuschätzen sind und begeben Sie sich gegebenenfalls auf die Suche nach weiteren Datensätzen!

Übungsfragen

1. Was unterscheidet Webscraping von der Datenerhebung über APIs?
2. Aus welchen Bestandteilen besteht eine URL?
3. Besuchen Sie eine Webseite und prüfen Sie, ob eine API bereit gestellt wird!
4. Schauen Sie sich die Dokumentation des Endpunkts „users“ der Twitter-API an. Suchen Sie dort die Bestandteile der verwendeten URL heraus: Wie lauten Domain, Pfad und Parameter?
5. Was versteht man unter Open Data?
6. Suchen Sie einen im Internet zum Download angebotenen Datensatz und schätzen Sie die Qualität der Daten ein. Was spricht gegen die wissenschaftliche Verwendung der gefundenen Daten, was spricht dafür?
7. Worin unterscheiden sich die Datenbanken von wissenschaftlichen Repositorien und die Datenbanken von Social-Media-Plattformen?

[12] Siehe Amazon Mechanical Turk (2018; https://www.mturk.com/).

Weiterführende Literatur

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures.* Dissertation, University of California.

Jünger, J. (2018). Mapping the field of automated data collection on the web. Data types, collection approaches and their research logic. In C. M. Stützer, M. Welker & M. Egger (Hrsg.), *Computational social science in the age of big data. Concepts, methodologies, tools, and applications* (S. 104–130). Köln: Halem.

Russell, M. A. (2014). *Mining the social web. Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more.* (2. Aufl.). Sebastopol: O'Reilly.

Literatur

Abgeordnetenwatch.de. (2022). *Abgeordnetenwatch API Dokumentation*. Zugriff am 30.05.2022. https://www.abgeordnetenwatch.de/api

Akademie der Wissenschaften und der Literatur Mainz. (2022). *Regesta Imperii*. Zugriff am 30.05.2022. http://www.regesta-imperii.de/

Amazon Mechanical Turk. (2018). *Amazon Mechanical Turk. Access a global, on-demand, 24x7 workforce*. Zugriff am 23.05.2022. https://www.mturk.com/

Amazon Web Services. (2022a). *Amazon Rekognition. Automatisieren Sie Ihre Bild- und Videoanalyse mit Machine Learning*. Zugriff am 30.05.2022. https://aws.amazon.com/de/rekognition/

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J. et al. (2021). *OSD2F: An Open-Source Data Donation Framework*. https://doi.org/10.31235/osf.io/xjk6t

BDZV. (2022). *Bundesverband Digitalpublisher und Zeitungsverleger*. Zugriff am 18.05.2022. https://www.bdzv.de/

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, *41*(1), 161–164. https://doi.org/10.1515/zgl-2013-0009

Beißwenger, M., Fladrich, M., Imo, W. & Ziegler, E. (2020). Die Mobile Communication Database 2 (MoCoDa 2). In K. Marx, H. Lobin & A. Schmidt (Hrsg.), *Deutsch in Sozialen Medien* (S. 349–352). Berlin: de Gruyter. https://doi.org/10.1515/9783110679885-018

Berlin-Brandenburgische Akademie der Wissenschaften. (2022). *API (Schnittstellen zum DWDS)*. Zugriff am 30.05.2022. https://www.dwds.de/d/api

Berlind, D. (2022). *ProgrammableWeb*. Zugriff am 18.05.2022. https://www.programmableweb.com/

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W. et al. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, *2004*(2), 597–620. https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/

Bruns, A. (2019). After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Center for Open Science. (2022). *Open Science Foundation*. Zugriff am 18.05.2022. https://osf.io/

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J. & Danescu-Niculescu-Mizil, C. (2020). *ConvoKit: A Toolkit for the Analysis of Conversations*. Proceedings of SIGDIAL. Zugriff am 30.05.2022. https://github.com/CornellNLP/ConvoKit

CKCEST: China Knowledge Centre for Engineering Sciences and Technology. (2022). *AMiner. Datasets for Social Network Analysis*. Zugriff am 30.05.2022. https://cn.aminer.org/data-sna

CLARIN. (2022). *CLARIN Virtual Language Observatory*. Zugriff am 21.06.2022. https://vlo.clarin.eu/

Datasette. (2022). *Openregister. Custom SQL query*. Zugriff am 30.05.2022. https://db.offeneregister.de/openregister

Destatis. (2022). *Statistisches Bundesamt*. Zugriff am 16.05.2022. https://www.destatis.de/

Dua, D. & Graff, C. (2019). *UCI Machine Learning Repository,* University of California, School of Information and Computer Science. Zugriff am 30.05.2022. https://archive.ics.uci.edu

Entity Reconciliation Community Group. (2022). *Reconciliation Service API v0.1. A protocol for data matching on the Web*. Zugriff am 16.05.2022. https://reconciliation-api.github.io/specs/0.1/

Evolution Institute & Seshat Project. (2021). *Seshat: Global History Databank*. Zugriff am 30.05.2022. http://seshatdatabank.info/

Fandom. (2022). *Fandom. Search the world's largest fan wiki platform*. Zugriff am 16.05.2022. https://www.fandom.com/

Fielding, R. (2000). *Architectural styles and the design of network-based software architectures*. Dissertation. Irvine: University of California.

Fiesler, C., Beard, N. & Keegan, B. C. (2020). No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 187–196. https://ojs.aaai.org/index.php/ICWSM/article/view/7290

Frey, J.-C. & Glaznieks, Aivars and Stemle, Egon W. (2019). *DIDI – The DiDi Corpus of South Tyrolean CMC 1.0.0,* Eurac Research CLARIN Centre. Zugriff am 30.05.2022. https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/7

Fu, K. & Zhu, Y. (2020). Did the world overlook the media's early warning of COVID-19? *Journal of Risk Research*, *23*(7–8), 1047–1051. https://doi.org/10.1080/13669877.2020.1756380

GermEval. (2019). *GermEval Shared Task Hub. Natural Language Processing shared tasks for German*. Zugriff am 30.05.2022. https://germeval.github.io/

Google. (2014). *Public Data*. Zugriff am 30.05.2022. https://www.google.com/publicdata/directory

Google. (2022d). *Cloud Vision API. Vision AI*. Zugriff am 30.05.2022. https://cloud.google.com/vision

Google. (2022f). *Dataset Search*. Zugriff am 30.05.2022. https://datasetsearch.research.google.com/

Google Developers. (2022). *YouTube. Let users watch, find, and manage YouTube content*. Zugriff am 30.05.2022. https://developers.google.com/youtube/

GovData. (2022). *Das Datenportal für Deutschland. Open Government: Verwaltungsdaten transparent, offen und frei nutzbar*. Zugriff am 30.05.2022. https://www.govdata.de/

Harvard University. (2022a). *Harvard Dataverse*. Zugriff am 18.05.2022. https://dataverse.harvard.edu/

Harvard University. (2022b). *Social Science One,* Harvard's Institute for Quantitative Social Science. Zugriff am 23.05.2022. https://socialscience.one/

Huynh, D. (2022). OpenRefine (Version 3.5.2) [Computer software]: Metaweb Technologies. https://openrefine.org/

Hydra W3C Community Group. (2018). *API Documentation*. Zugriff am 18.05.2021. http://www.hydra-cg.com/drafts/use-cases/2.api-documentation.md

IBM. (2022). *IBM Watson Studio. Erstellen und skalieren Sie vertrauenswürdige KI in jeder Cloud. Automatisieren Sie den KI-Lebenszyklus für ModelOps.* Zugriff am 30.05.2022. https://www.ibm.com/de-de/cloud/watson-studio

International Movie Database. (2022). *IMDb Datasets*. Zugriff am 23.05.2022. https://www.imdb.com/interfaces/

IVW. (2022). *Informationsgesellschaft zur Feststellung der Verbreitung von Werbeträgern*. Zugriff am 16.05.2022. https://www.ivw.de/

Jacobson, D., Brail, G. & Woods, D. (2012). *APIs: A strategy guide. Creating channels with application programming interfaces.* Beijing: O'Reilly.

Johnson, B. & Turner, L. A. (2003). Data Collection Strategies in Mixed Methods Research. In A. Tashakkori & C. Teddlie (Hrsg.), *Handbook of Mixed Methods in Social & Behavioral Research* (S. 297–319). Thousand Oaks: Sage.

Jünger, J. (2021). A brief history of APIs. How social media providers shape the opportunities and limitations of online research. In U. Engel, A. Quan-Haase, S. X. Liu & L. Lyberg (Hrsg.), *Handbook of Computational Social Science* (S. 17–32). London: Routledge. https://doi.org/10.4324/9781003025245-3

Kaggle. (2022). *Kaggle: Your Machine Learning and Data Science Community*. Zugriff am 23.05.2022. https://www.kaggle.com/

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J. et al. (2016). D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. *PloS one*, *11*(7), e0158391. https://doi.org/10.1371/journal.pone.0158391

Kotsios, A., Magnani, M., Vega, D., Rossi, L. & Shklovski, I. (2019). An Analysis of the Consequences of the General Data Protection Regulation on Social Network Research. *ACM Transactions on Social Computing*, *2*(3), 1–22. https://doi.org/10.1145/3365524

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf

Leetaru, K. H. (2021). *The GDELT Project*. Zugriff am 08.05.2022. https://www.gdeltproject.org/

GESIS. (2022). *GESIS-Suche*. Zugriff am 18.05.2022. https://search.gesis.org/

Linvill, D. & Warren, P. (2018). *3 million Russian troll tweets*. Zugriff am 30.05.2022. https://github.com/fivethirtyeight/russian-troll-tweets/

Media Cloud (2022). *Media Cloud is an open-source platform for media analysis*. Zugriff am 30.05.2022. https://mediacloud.org/

MediaWiki. (2022, 16. Mai). *API:Main page*. Zugriff am 30.05.2022. https://www.mediawiki.org/wiki/API:Main_page

Meta. (2021). *CrowdTangle. A tool from Meta to help follow, analyze, and report on what's happening across social media.* Zugriff am 23.05.2022. https://www.crowdtangle.com/
Meta. (2022). *Meta for Developers*. Zugriff am 30.05.2022. https://developers.facebook.com/
Microsoft. (2022a). *Azure Cognitive Services documentation. Learn how to use ready-made AI services to build intelligent apps, websites, and bots. Develop software that can see, hear, speak, and interpret your user's needs.* Zugriff am 30.05.2022. https://docs.microsoft.com/en-us/azure/cognitive-services/
Ministerium der Justiz Nordrhein-Westfalen. (2020). *Gemeinsames Registerportal der Länder*. Zugriff am 16.05.2022. https://www.handelsregister.de/rp_web/welcome.xhtml
NFDI. (2022). *Nationale Forschungsdaten Infrastruktur*. Zugriff am 15.01.2023. https://www.nfdi.de
OpenAPI Initiative (2022). *The OpenAPI Specification Repository*. Zugriff am 16.05.2022. https://github.com/OAI/OpenAPI-Specification
OpenCorporates. (2022). *Get data access to over 200 million companies*. Zugriff am 30.05.2022. https://api.opencorporates.com/
OpenRefine. (2021). *Reconcilable Data Sources*. Zugriff am 18.05.2022. https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources
Peroni, S. & Daquino, M. (2020). *The unifying REST API for all the OpenCitations Indexes*. Zugriff am 30.05.2022. https://opencitations.net/index/api/v1
Perspective. (2021). *Using machine learning to reduce toxicity online. Perspective API can help mitigate toxicity and ensure healthy dialogue online*. Jigsaw. Zugriff am 30.05.2022. https://www.perspectiveapi.com/
Puschmann, C. & Haim, M. (2021). *useNews*. [Dataset]. Zugriff am 28.01.2022. https://osf.io/uzca3/
RatSWD. (2019). *Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement.* Berlin: RatSWD. https://doi.org/10.17620/02671.39
Reese, A. (2020). *Trump Tweets. Tweets from @realdonaldtrump*. [Dataset], kaggle.com. Zugriff am 30.05.2022. https://www.kaggle.com/datasets/austinreese/trump-tweets
Rossi, L. & Ahmed, N. K. (2022). *Network Repository. A Scientific Network Data Repository with Interactive Visualization and Mining Tools*. Zugriff am 30.05.2022. https://networkrepository.com/
Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. (2005). Effects of Age and Gender on Blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. http://www.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf
SemEval. (2022). *International Workshop on Semantic Evaluation*. Zugriff am 30.05.2022. https://semeval.github.io/
Thelwall, M. [Mike] & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, *57*(13), 1771–1779. https://doi.org/10.1002/asi.20388
Twitter. (2022d). *Twitter Developer Platform. Use Cases, Tutorials, & Documentation*. Zugriff am 30.05.2022. https://developer.twitter.com/en
University of Michigan. (2022). *ICPSR: Inter-University Consortium for Political and Social Research*. Zugriff am 30.05.2022. https://www.icpsr.umich.edu/web/pages/
University of Oxford. (2021). *OTA: Oxford Text Archive. A repository of full-text literary and linguistic resources. Thousands of texts in more than 25 languages*. Zugriff am 30.05.2022. https://ota.bodleian.ox.ac.uk/repository/xmlui/

Uppsala Conflict Data Program. (2016). *Georeferenced Event Dataset (GED). Global version 17.1*. [Dataset]. http://ucdp.uu.se/downloads/

Visual Geometry Group. (2022). *VoxCeleb. A large scale audio-visual dataset of human speech*. Zugriff am 30.05.2022. https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

Weiboscope. (2022). *HKU JMSC Weibo Censorship Index*. Journalism and Media Studies Centre. Zugriff am 23.05.2022. https://weiboscope.jmsc.hku.hk/wsr/

Wikimedia Deutschland. (2022). *Wikipedia. Die freie Enzyklopädie*. Zugriff am 16.05.2022. https://www.wikipedia.de/

Wikipedia. (2022a). *List of datasets for machine-learning research*. https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 1–9. https://doi.org/10.1038/sdata.2016.18

Zubiaga, A. (2018a). A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, *69*(8), 974–984. https://doi.org/10.1002/asi.24026

Zubiaga, A. (2018b). *Twitter event datasets 2012–2016*. [Dataset]. Zugriff am 30.05.2022. https://doi.org/10.6084/m9.figshare.5100460.v2