Rat für Sozial- und Wirtschaftsdaten (RatSWD) (Ed.)

Research Report

Erhebung und Nutzung unstrukturierter Daten in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Herausforderungen und Empfehlungen

RatSWD Output Series, 7. Berufungsperiode, No. 2

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Rat für Sozial- und Wirtschaftsdaten (RatSWD) (Ed.) (2023) : Erhebung und Nutzung unstrukturierter Daten in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Herausforderungen und Empfehlungen, RatSWD Output Series, 7. Berufungsperiode, No. 2, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin, https://doi.org/10.17620/02671.73

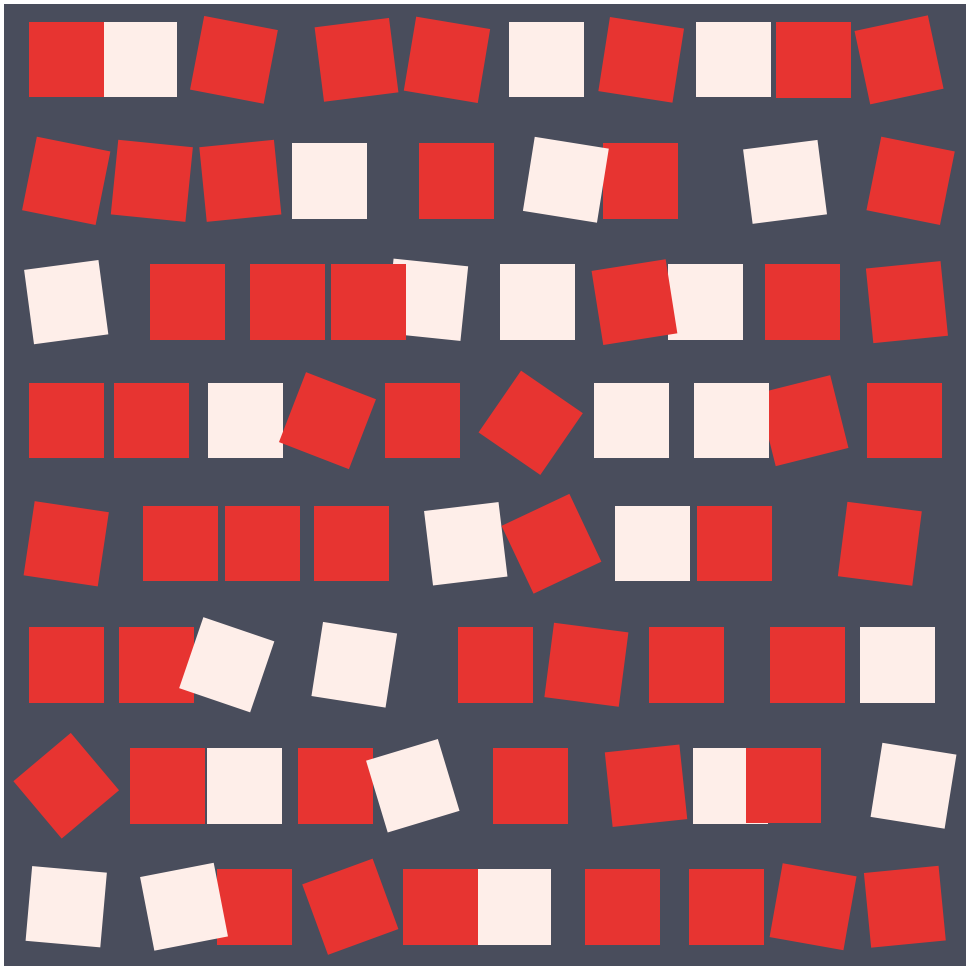
This Version is available at:
https://hdl.handle.net/10419/270310

2 | Output
7. Berufungsperiode (2020-2023)

RatSWD.
Rat für Sozial- und WirtschaftsDaten

Erhebung und Nutzung **unstrukturierter Daten** in den Sozial-, Verhaltens- und Wirtschaftswissenschaften

Herausforderungen und Empfehlungen



Rat für Sozial- und Wirtschaftsdaten (RatSWD)

Erhebung und Nutzung **unstrukturierter Daten** in den Sozial-, Verhaltens- und Wirtschaftswissenschaften

Herausforderungen und Empfehlungen

Abkürzungsverzeichnis

API . . . . . Application Programming Interface
DGPuK . . . . . Deutsche Gesellschaft für Publizistik und Kommunikationswissenschaft
EDMO . . . . . European Digital Media Observatory
GPS. . . . . Global Positioning System
HTML . . . . . Hypertext Markup Language
NFDI. . . . . Nationale Forschungsdateninfrastruktur
TEF . . . . . Total Error Framework
TSE framework . . . . . Total Survey Error Framework
URL. . . . . Uniform Resource Locator
XML . . . . . Extensible Markup Language

Inhaltsverzeichnis

Abbildungsverzeichnis

Abstract

Die zunehmende Digitalisierung unserer Lebenswelt in den letzten Jahrzehnten hat zu einer Reihe von neuen Datenquellen für die Sozial-, Verhaltens- und Wirtschaftswissenschaften geführt. Hierzu gehören vor allem auch unstrukturierte Daten, die sich dadurch auszeichnen, dass sie nicht in Form eines festen Datenformats vorliegen und daher nicht einfach datenanalytisch weiterverarbeitet werden können (z. B. Facebook-Texte, Instagram-Bilder, YouTube-Videos, Twitter-Nachrichten). Die Nutzung unstrukturierter Daten ist mit spezifischen Herausforderungen verknüpft, die gerade dadurch entstehen, dass die Daten typischerweise nicht in einer kontrollierten wissenschaftlichen Studie erhoben werden, sondern häufig im natürlichen Lebensumfeld anfallen. Aufbauend auf den Ergebnissen eines Expert:innen-Workshops werden die spezifischen Herausforderungen bei der Erhebung und Nutzung unstrukturierter Daten beschrieben und Empfehlungen formuliert. Diese orientieren sich am Total Error Framework und beziehen sich auf die Datengenerierung (Definition von Untersuchungseinheiten, Coverage und Sampling Error, Nonresponse und Missing Data Error), die Datenaufbereitung (Spezifikationsfehler, Validität, Messfehler und inhaltliche Fehler) sowie die Datenanalyse (Record Linkage und Verarbeitungsfehler, Modellierungsfehler, analytische Fehler). Abschließend werden offene Fragen und Herausforderungen bei der Forschung mit unstrukturierten Daten diskutiert. Der Output richtet sich einerseits an Studierende sowie Forschende der Sozial-, Verhaltens- und Wirtschaftswissenschaften, andererseits an alle, die mit unstrukturierten Daten arbeiten und Schlüsse aus diesen für praktische Anwendungsfragen ziehen.

1 Einleitung

1.1 Definition von unstrukturierten Daten und Abgrenzung zu anderen Begriffen

Die zunehmende Digitalisierung unserer Lebenswelt in den letzten Jahrzehnten hat zu einer Reihe von neuen Datenquellen für die Sozial-, Verhaltens- und Wirtschaftswissenschaften geführt. Diese Daten haben häufig gemeinsam, dass sie nicht in Form eines festen Datenformats vorliegen (z. B. eines rechteckigen Schemas mit Fällen/Beobachtungen in Zeilen und Variablen in Spalten) und daher nicht einfach datenanalytisch weiterverarbeitet werden können (Eberendu, 2016). Im Folgenden sollen diese Datenquellen unter dem Begriff ***unstrukturierte Daten*** zusammengefasst werden und von den traditionellen Survey-Daten abgegrenzt werden. Beispiele für unstrukturierte Daten sind in Abbildung 1 zusammengestellt. Unstrukturierte Daten zeichnen sich häufig durch ein hohes Volumen aus und erfordern dann umfangreiche Aufbereitungen, um sie der sozial-, verhaltens- und wirtschaftswissenschaftlichen Forschung zugänglich zu machen. Strukturierte Daten liegen hingegen in Form eines festen Datenformats (z. B. Tabellen, Datensätze, Datenbanken) vor (Tanwar et al., 2015). Eine Zwischenstellung nehmen semistrukturierte Daten ein. Diese liegen nicht in einem festen Format wie strukturierte Daten vor. Sie können aber Strukturelemente enthalten und einfach ausgetauscht werden (Eberendu, 2016; Tanwar et al., 2015). Ein Beispiel sind Extensible-Markup-Language-Daten (XML-Daten; Tanwar et al., 2015). XML-Daten weisen eine partielle (z. B. hierarchische) Struktur auf und enthalten Strukturmerkmale wie z. B. Auszeichnungen (*Tags*; Nyhuis, 2021). Die aus dem Internet bekannten Hypertext-Markup-Language-Dokumente (HTML-Dokumente) sind beispielsweise spezielle XML-Dokumente (Bosse et al., 2021).

Abbildung 1: Ausgewählte Beispiele für unstrukturierte Daten

Soziale Medien	Facebook-Texte, Instagram-Bilder, YouTube-Videos, Twitter-Nachrichten
Allgemeine Medien	Texte, Bilder, Videos, Sprachaufzeichnungen, Musik
Geodaten	GPS-Daten
Log-Daten	Besuch von Webseiten, Verweildauer auf Webseiten, E-Mail-Verhalten
World Wide Web	Webseiten, Nachrichten, Blogs
Allgemeine Dokumente	Texte, PDF-Dateien, eingescannte Dateien
Finanzdaten	Banktransaktionen, Börsendaten
Gesundheitsdaten	Patientenakten, Röntgenbilder, Scanner-Bilder

Quelle: in Anlehnung an Eberendu (2016) sowie Taleb et al. (2018)

Unstrukturierte Daten weisen Ähnlichkeiten zu anderen Datentypen auf. So können unstrukturierte Daten *Big Data* untergeordnet werden, wobei der Begriff *Big Data* nicht eindeutig definiert ist. Unter ***Big Data*** werden typischerweise Daten verstanden, die sich durch einen hohen Datenumfang und eine hohe Vielfalt auszeichnen und mit hoher Geschwindigkeit erzeugt werden (Gandomi & Haider, 2015; Lazer & Radford, 2017; Tanwar et al., 2015). Eine hohe Vielfalt zeichnet sich durch die strukturelle Heterogenität eines Datensatzes aus und den Umstand, dass sowohl strukturierte als auch semistrukturierte und unstrukturierte Daten erhoben werden (Gandomi & Haider, 2015).

Unstrukturierte Daten umfassen häufig Daten, die anhand ***neuer Informationstechnologie*** (RatSWD, 2020) z. B. im Internet oder anhand von Smartphones gewonnen werden und das digitale Leben von Menschen (z. B. Facebook- oder Twitter-Daten) kennzeichnen (Lazer & Radford, 2017). Sie können auch Aspekte des sogenannten *digitalisierten* Lebens erfassen, das nach Lazer und Radford (2017) Aspekte digitalen Lebens repräsentiert, die nicht von einer Person aktiv produziert werden (wie z. B. Tweets), sondern durch die Digitalisierung anfallen (z. B. die Erfassung sozialer Nähe via Bluetooth). Die Unterscheidung in digitales vs. digitalisiertes Leben korrespondiert mit der Unterscheidung in intentional

vs. nicht-intentional anfallende Daten (Hox, 2017) bzw. in Partizipationsspuren (aktives Verhalten) vs. Transaktionsdaten (z. B. Metadaten über das digitale Verhalten wie der Ort; Menchen-Trevino, 2013). Schließlich können auch digitale Spuren (anfallende Daten) aufgezeichnet werden, die die Nutzung digitaler Geräte hinterlassen. Lazer und Radford (2017) verweisen diesbezüglich z. B. auf Merkmale, die die Telefonnutzung beschreiben. Unstrukturierte Daten sind aber nicht mit Daten gleichzusetzen, die anhand neuer Informationstechnologie gewonnnen werden, da letztere auch auf anderem Wege gewonnen werden können. Hier wären vor allem aufwändige Textdaten zu nennen (Grimmer et al., 2022).

Unstrukturierte Daten sind häufig Daten, die aufgrund ***nicht-reaktiver*** Erhebung gewonnen werden. Datenerhebungen gelten als reaktiv, wenn ihre Ausprägungen durch die Art der Datenerhebung bedingt von Untersuchungsteilnehmenden und/oder Untersuchenden beeinflusst werden können (Fritsche & Linneweber, 2006). Unstrukturierte Daten sind häufig Daten, die anfallen, ohne dass den Datengebenden bewusst ist, dass sie an einer wissenschaftlichen Studie teilnehmen oder Daten für diese Zwecke genutzt werden (*unobtrusive measures*; Webb et al., 1966). Untersuchungsbedingte Datenbeeinflussungen sind daher nicht möglich oder zumindest unwahrscheinlich. Unstrukturierte Daten sind aber nicht mit nicht-reaktiv erhobenen Daten gleichzusetzen, da es auch strukturierte nicht-reaktive Messungen (z. B. strukturierte Verhaltensbeobachtungen) gibt und Daten auch in Kontexten erzeugt werden, in denen Personen bewusst ist, dass ihre Daten weiterverwendet werden können (z. B. Nutzung von Suchmaschinen). In den Sozialwissenschaften ist die Unterscheidung in *found data* und *designed data* geläufig (Biemer & Amaya, 2020). Unter ***found data*** versteht man Daten, die nicht primär für eine wissenschaftliche Studie erhoben wurden, sondern vorgefunden werden (z. B. Archivdaten), während Survey-Daten ein Beispiel für ***designed data*** sind, da sie für wissenschaftliche Zwecke konzipiert und erhoben wurden. Solche vorgefundenen Daten sind häufig unstrukturierte Daten, die für die statistische Analyse erst aufbereitet werden müssen.

Unstrukturierte Daten sind typischerweise Daten, die im ***natürlichen Lebensumfeld*** (naturalistische Daten, Feldforschung) und nicht im Labor entstehen. Sie sind aber nicht mit naturalistischen Daten identisch, da letztere sich auch auf strukturierte Daten beziehen können (z. B. Erfassung der Stimmung mittels *Ambulatory Assessment*).

1.2 Bedeutung von unstrukturierten Daten

Unstrukturierte Daten sind für die Sozial-, Verhaltens- und Wirtschaftswissenschaften aus unterschiedlichen Gründen von großer Bedeutung. So fallen unstrukturierte Daten in sehr vielen Lebensbereichen an, sie bilden wichtige Teile menschlichen Lebens ab, die über strukturierte Daten in der Form nicht abgebildet werden können. Ein Großteil dieser anfallenden Daten wird aber noch nicht ausgewertet (Eberendu, 2016).

Wie bereits im einleitenden Abschnitt dargelegt wurde, handelt es sich bei unstrukturierten Daten häufig um Daten, die nicht-reaktiv im natürlichen Lebensumfeld erhoben werden bzw. anfallen. Hierdurch werden Verzerrungen, die bei reaktiv oder im Labor erhobenen Daten wirksam werden können, vermieden. Andererseits können natürlich auftretende Phänomene mit hoher ökologischer Validität erfasst werden. Sie haben das Potenzial, in hohem Maße die Lebenswirklichkeit in der natürlichen Umwelt, sozialen Gruppen und Organisationen abzubilden.

Da unstrukturierte Daten häufig das konkrete Verhalten widerspiegeln, umgehen sie zum Teil Probleme, die mit dem Selbstbericht in klassischen Survey-Studien verknüpft sind (z. B. Antwortstile, soziale Erwünschtheit, Selbst- und Fremdtäuschung; Borkenau, 2006). Sie können traditionelle Survey-Studien daher auch in sinnvoller Weise ergänzen, indem sie Möglichkeiten der Validierung von Survey-Daten bieten (Jürgens et al., 2020) sowie Survey-Daten zur Erklärung sozialer Phänomene durch andere Datenquellen (z. B. Verhaltensdaten) anreichern und somit die Erklärungskraft von Survey-Studien erhöhen (Reveilhac et al., 2022).

1.3 Ziele und Adressat:innen des Outputs

Unstrukturierte Daten eröffnen der sozial-, verhaltens- und wirtschaftswissenschaftlichen Forschung neue Perspektiven. Die Nutzung unstrukturierter Daten ist aber auch mit spezifischen Herausforderungen verknüpft, die gerade dadurch entstehen, dass die Daten typischerweise nicht in einer kontrollierten wissenschaftlichen Studie erhoben werden. Diese besonderen Herausforderungen sollen im Folgenden beleuchtet werden und – sofern möglich – Empfehlungen zum Umgang mit diesen Herausforderungen formuliert werden.

Der Output richtet sich einerseits an Studierende sowie Forschende der Sozial-, Verhaltens- und Wirtschaftswissenschaften, andererseits an alle, die mit unstrukturierten Daten arbeiten und Schlüsse aus diesen für praktische Anwendungsfragen ziehen.

1.4 Kurzer Bericht zur Befragung und Workshop

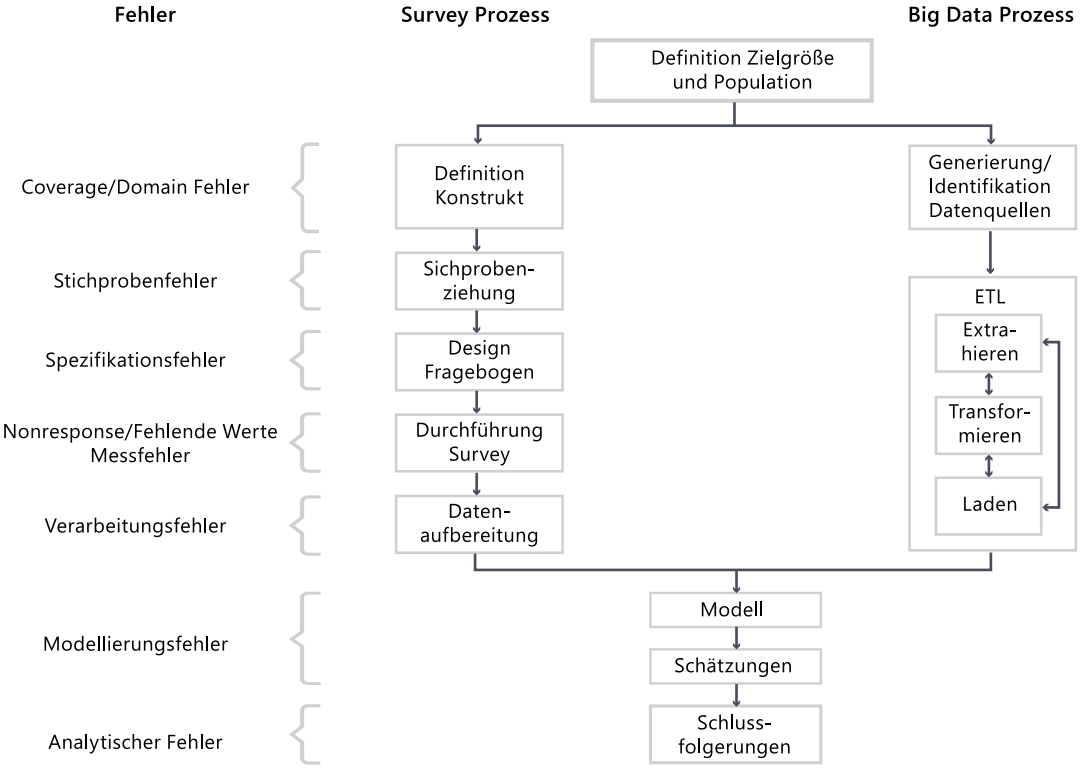
Im Zuge dieser Output-Erstellung wurde auf Grundlage des *Total Error Frameworks* (TEF) für *Big Data* (Amaya et al., 2020) ein Fragebogen zur Qualitätssicherung unstrukturierter Daten entwickelt. Dieser besteht aus insgesamt 32 Fragen und umfasst neben Fragen zu allgemeinen fachlichen Angaben der Teilnehmenden auch Fragen zur Generierung, Aufbereitung und Analyse von unstrukturierten Daten sowie eine offene Abschlussfrage. An der Befragung nahmen 19 Wissenschaftler:innen aus verschiedenen relevanten Wissenschaftsgebieten teil (Betriebswirtschaftslehre, Bildungsforschung, Computational Social Science, Kommunikationswissenschaft, Politikwissenschaft, Psychologie, Soziologie und Volkswirtschaftslehre). In ihrer Forschung verwenden sie ein breites Spektrum von unterschiedlichen Datentypen: Social-Media-Daten (z. B. Twitter, Facebook), Smartphone-Daten, Audiodaten, klassische Medien (z. B. Zeitungen) und Textdokumente (z. B. Parlamentsreden, Betriebsvereinbarungen, Abrechnungen).

Die Hauptergebnisse der Befragung wurden in einem zweitägigen Online-Workshop am 13. und 14. Oktober 2021 mit den Wissenschaftler:innen diskutiert. Der Workshop war in vier thematische Sessions unterteilt: Datengenerierung, Datenaufbereitung, Datenanalyse und offene Fragen zur Verfügbarkeit neuer Datentypen. Der vorliegende Output fasst die Ergebnisse der Befragung und Diskussion zusammen und bettet sie in das TEF ein.

1.5 Kurze Einführung in Total Error Frameworks zur Beurteilung von Datenqualität

Total-Error-Framework-Ansätze entstehen aktuell in verschiedenen Versionen und mit Fokus auf verschiedene Datenarten. Für die Diskussion der Qualitätssicherung unstrukturierter Daten orientieren wir uns konkret am ***TEF von Amaya et al. (2020)***. Dieses TEF stellt eine kürzlich vorgeschlagene Übertragung des in der Survey-Forschung verbreiteten Error-Framework-Prinzips auf *Big Data* dar, das insbesondere in Form des *Total Survey Error Frameworks (TSE framework;* Groves & Lyberg, 2010) etabliert ist. Mit dem *TSE framework* werden alle Fehlerquellen zusammengefasst, die einen möglichen Einfluss auf die Ergebnisse einer empirischen, umfragebasierten Untersuchung besitzen. Insgesamt berücksichtigt das TEF acht Fehlerquellen, mit denen der gesamte Prozess aus Datengenerierung, Datenaufbereitung und Datenauswertung abgedeckt werden soll. Abbildung 2 zeigt das TEF (Amaya et al., 2020), wobei die Boxen auf der linken Seite (Survey-Prozess) bzw. der rechten Seite (Big-Data-Prozess) jeweils Schritte hervorheben, die spezifisch für die Survey-Forschung bzw. den Umgang mit *Big Data* sind. Der mittlere Teil ist beiden Ansätzen gemein und unabhängig von der Datenquelle bzw. Struktur der Daten. Das TEF bietet eine hilfreiche erste Strukturierung für die Diskussion der methodischen Herausforderungen bei der wissenschaftlichen Arbeit mit unstrukturierten Daten. Der vorliegende Bericht ist in die Kapitel „Datengenerierung", „Datenaufbereitung" und „Datenanalyse" gegliedert und orientiert sich an den in Abbildung 2 thematisierten Herausforderungen.

Abbildung 2: Total Error Framework (TEF)



Quelle: Amaya et al. (2020)

Das Kapitel **Datengenerierung** thematisiert die Definition der Untersuchungseinheiten sowie damit verbundene *Coverage Error*, die dadurch entstehen, dass Zielpopulation und Erhebungspopulation nicht deckungsgleich sind. Des Weiteren werden klassische Stichprobenfehler angesprochen und von Fehlern unterschieden, die durch einen Stichprobenausfall (*Nonresponse*) auf Ebene der Einheiten (*Unit Nonresponse*) oder innerhalb einer Einheit (*Item Nonresponse*) entstehen. Das Kapitel **Datenaufbereitung** beinhaltet mit dem Spezifikationsfehler mögliche Abweichungen zwischen dem aufgrund der inhaltlichen Fragestellung zu erfassendem Konzept und dem empirischen gemessenen Konzept. Diese Abweichungen umfassen sowohl Einschränkungen bezüglich der verfügbaren Daten als auch Verfälschungen, die durch technische Fehler oder individuelle Einstellungen von Geräten entstehen können. In den Sozialwissenschaften und der Psychometrie werden diese Aspekte auch unter den Begriffen der Validität und Reliabilität von Messungen zusammengefasst (Eid & Schmidt, 2014). Im Kapitel **Datenanalyse** werden mit dem Verarbeitungsfehler (*Processing Error*) mögliche Verzerrungen angesprochen, die auf die Eingabe, Transformation oder Kodierung von Variablen zurückzuführen sind. Verarbeitungsfehler können aber auch durch die Verknüpfung von verschiedenen Datenquellen für eine Untersuchungseinheit (*Record Linkage*) entstehen. Weitere Fehlerquellen in der Datenanalyse umfassen die Behandlung von fehlenden Daten in den statistischen Analysen (Modellierungsfehler) sowie Verzerrungen, die durch eine fehlerhafte Anwendung von statistischen Modellen und der Interpretation der Analyseergebnisse entstehen (Analytischer Fehler). Nach dem TEF liegen die besonderen Herausforderungen bei der Arbeit mit unstrukturierten Daten vor allem in den Bereichen **Datengenerierung** und **Datenaufbereitung** (siehe Abbildung 2), wohingegen die möglichen Fehlerquellen bei der Datenanalyse von ähnlicher Natur sind wie bei der traditionellen Survey-Forschung.

Es sei abschließend betont, dass der vorliegende Bericht lediglich als ein erster Schritt zur Verständigung über Qualitätsstandards für die Arbeit mit unstrukturierten Daten angesehen werden sollte. Das primäre Ziel besteht darin, besondere Herausforderungen herauszuarbeiten, die sich im Vergleich zur Arbeit mit Daten aus der traditionellen Survey-Forschung stellen. Das TEF bietet hierfür eine hilfreiche Orientierung in einem sich schnell entwickelnden Forschungsfeld. Die Ausformulierung von expliziten Standards muss dann vor dem Hintergrund konkreter Datentypen vorgenommen werden. Aktuell wird die Grundidee der *Error Frameworks* bereits weiter ausdifferenziert und auch auf verschiedene Datenarten angewandt (z. B. Twitter-Daten, siehe Hsieh & Murphy, 2017; Daten aus Online-Plattformen, siehe Sen et al., 2021; *Metered Data*, siehe Bosch & Revilla, 2022).

2 Datengenerierung

Dieses Kapitel klärt über die Besonderheiten der ***Datengenerierung*** auf und präsentiert sowohl die speziellen Problemlagen und Herausforderungen als auch Empfehlungen zum Umgang mit unstrukturierten Daten im Generierungsprozess. Dabei wird zunächst die *Definition der Untersuchungseinheiten*, anschließend der *Coverage Error* und *Sampling Error* sowie abschließend der *Nonresponse* bzw. *Missing Data Error* in Hinblick auf unstrukturierte Daten diskutiert.

2.1 Definition von Untersuchungseinheiten und Datenstruktur

Bei unstrukturierten Daten können ***sehr unterschiedliche Untersuchungseinheiten*** gewählt werden. Selbst innerhalb von Social-Media-Daten können beispielsweise das Individuum, der Account, der Inhalt eines Beitrags (z. B. „Post" oder Kommentar), oder die Interaktion (z. B. Follower-Netzwerke) als Untersuchungseinheit definiert werden. Des Weiteren können ganze Texte in Archiven als Untersuchungseinheit gewählt werden oder einzelne Sequenzen von Bild- oder Tonmaterial. Die Definition der Untersuchungseinheit hat gerade bei unstrukturierten Daten ***weitreichende Folgen*** für die Klassifizierung diverser Fehlerquellen. Mit der Definition der Untersuchungseinheit bzw. der Grundgesamtheit – beispielsweise Personen im Allgemeinen vs. User/Accounts – bemessen sich auch *Coverage Error* und *Sampling Error*, die im nächsten Abschnitt diskutiert werden.

Häufig weisen die Daten eine ***Mehrebenenstruktur*** auf. So sind etwa Tweets geclustert in Accounts (z. B. Fischer et al., 2019), Beobachtungszeitpunkte in Personen oder Personen in räumlichen Aggregaten. Fehlerfreie Informationen zur Struktur der Clusterung sind je nach Definition der Untersuchungseinheiten auf den jeweiligen Ebenen nicht immer gegeben. So können möglicherweise Tweets zu Accounts korrekt zugeordnet werden, jedoch nicht zwingend zu Personen. Eine Person kann mehrere Accounts nutzen (z. B. unterschiedliche Facebook-Accounts eines Jugendlichen für Freunde und Eltern/Familie) oder mehrere Personen können einen Facebook-Account bedienen (z. B. im gewerblichen Kontext). Die teils unklare Datenstruktur bzw. ungenaue Zuordnung der Untersuchungseinheiten auf Ebene 1 zu der übergeordneten Ebene 2 wirkt sich spätestens bei der Wahl einer adäquaten Modellierung, die die Verletzung der Unabhängigkeitsannahme der Fälle berücksichtigen sollte (etwa durch die Anwendung von Mehrebenen-/Paneldatenanalysen oder robusten Standardfehlern), nachteilig aus (siehe auch Kapitel 4).

2.2 Coverage Error und Sampling Error

Im *TSE framework* sind der *Coverage Error* und der *Sampling Error* klar unterschieden. Der ***Coverage Error*** verweist auf Differenzen zwischen der Grundgesamtheit, auf die die Ergebnisse der Forschung verallgemeinert werden sollen, und dem Stichprobenplan, einer Liste oder einem Verfahren zur Aufzählung aller Elemente der Grundgesamtheit. Der ***Sampling Error*** verweist auf Differenzen zwischen dem Stichprobenplan und der daraus gezogenen Stichprobe, die zufällig (Stichprobenfehler) oder auch systematisch auftreten können. Dass diese Unterscheidung für unstrukturierte Daten in sehr vielen Fällen nicht getroffen werden kann, legten sowohl die Expert:innen-Befragung als auch der Expert:innen-Workshop nahe, die im Rahmen dieser Arbeit durchgeführt wurden. Das kann unterschiedliche Ursachen haben: In vielen Fällen sind der Stichprobenplan und die Stichprobe identisch, weil – anders als bei Surveys – keine wirtschaftliche Notwendigkeit besteht, die Zahl der Untersuchungseinheiten durch Ziehung einer Stichprobe einzuschränken. Andererseits gibt es auch Fälle, in denen Datenanbieter direkt Stichproben über eine *Application Programming Interface* (API) vertreiben und somit für Forschende gar kein Zugang zum Stichprobenplan besteht, sodass dieser sich der separaten Untersuchbarkeit entzieht. Aus diesen Gründen werden im Folgenden Selektionen auf Ebene des Stichprobenplans und der Stichprobe gemeinsam behandelt.

Zu unterscheiden sind Studien, in denen Daten, z. B. Accounts, von Plattformen ***ohne Zustimmung*** der Urheber:innen extrahiert werden können (Typ A), von Studien, die nur ***mit Zustimmung*** der Teilnehmenden und ggf. einer umfangreichen Mitwirkung (z. B. Installation einer Smartphone-App) und Nutzung eigener Hardware (z. B. Wearables) Daten erheben können (Typ B).

Bei Typ A (Studien ***ohne*** Zustimmung der Teilnehmenden) werden insbesondere die folgenden ***Probleme und Herausforderungen*** identifiziert:

1. ***Anpassung der Grundgesamtheit an die Datenverfügbarkeit („Verfügbarkeitsforschung"):*** Das Problem, dass die Nutzenden bestimmter Dienste (z. B. Facebook, Twitter) nicht die gesamte Bevölkerung (z. B. eines Landes) repräsentieren, kann man prinzipiell lösen[1], indem man die Nutzenden, für die Daten verfügbar sind, zur Grundgesamtheit erklärt. Es sollten dann aber auch keine Schlüsse auf größere Populationen gezogen werden, ohne das Selektionsproblem zu adressieren.
2. ***Unklare Algorithmen:*** Die Datenanbieter nutzen proprietäre und intransparente Algorithmen zur Auswahl von Stichproben, die der Forschung über APIs verfügbar gemacht werden. Diese können sowohl Auswirkungen auf die Auswahl als auch auf Sortierungen haben. So unterscheiden sich beispielsweise kostenlose 1-Prozent-Stichproben der Twitter-Daten von kommerziellen 10-Prozent-Stichproben (Morstatter et al., 2013). Forschende, die mit den Daten des Anbieters arbeiten möchten, haben häufig keine Alternative zur Nutzung der durch intransparente Algorithmen ausgewählten Daten.
3. ***Paywalls:*** Insbesondere bei der Analyse von Textdokumenten spielt die Tatsache eine größere Rolle, dass ein Teil der Dokumente nicht frei zugänglich ist und daher von Forschenden in vielen Fällen nicht genutzt werden kann. Es steht zu vermuten, dass sich diese Dokumente von den frei zugänglichen in vielen Merkmalen unterscheiden.
4. ***Personalisierte Darstellungen:*** Der Inhalt, der beim Besuch z. B. einer Webseite angezeigt wird, ist unter Umständen abhängig von Eigenschaften und vom vorherigen Verhalten der Nutzenden und somit nicht objektiv feststellbar. Dieses Problem strahlt stark auf die Replizierbarkeit von Forschung aus.
5. ***Löschung von Inhalten:*** Auf Plattformen werden regelmäßig Inhalte oder Accounts gelöscht, wenn sie z. B. Richtlinien verletzen oder Nutzende dies verlangen. Die Grundgesamtheit aller getätigten „Posts" auf einer Plattform ist daher unter Umständen schon nach kurzer Zeit nicht mehr verfügbar. Je weiter der zeitliche Abstand zwischen Beitrag und Extraktion der Stichprobe, desto mehr ist die Stichprobe dann verzerrt in Richtung regelkonformer Beiträge. Das stellt insbesondere dann ein massives Problem dar, wenn etwa Regelverletzungen Gegenstand der Forschung sind. In anderen Fällen kann technischer Wandel eine Ursache darstellen, dass Inhalte nicht weiter verfügbar sind (z. B. veraltete Datenformate). Auch dieses Problem strahlt stark auf die Replizierbarkeit von Forschung aus.
6. ***Overcoverage durch Bots (automatisierte Accounts):*** Ein erheblicher Teil der Accounts und Beiträge auf Online-Plattformen ist auf automatisierte Accounts, sogenannte Bots, zurückzuführen und somit nicht Teil der Grundgesamtheit der meisten Forschungsvorhaben. Werden Bots nicht als solche erkannt, kann das Ergebnisse erheblich verzerren.
7. ***Selektionsfehler bei Stichproben:*** Werden aus dem Gesamtbestand einer Plattform für den Forschungszweck Stichproben beispielsweise durch Auswahl von Beiträgen, die bestimmte Stichworte enthalten, gezogen, dann kann es sowohl auf der Ebene von Beiträgen als auch auf der Ebene von Nutzendengruppen zu Selektionsfehlern kommen. Dies würde z. B. zutreffen, wenn etwa bestimmte Altersgruppen bevorzugt bestimmte abweichende Begrifflichkeiten oder ethnische Gruppen andere Sprachen verwenden (Sen et al., 2021).
8. ***Datenschutz:*** Datenschutzrechtliche und forschungsethische Beschränkungen können ebenfalls eine Ursache dafür sein, dass selektiv auf einen Teil eines Datenbestands nicht zugegriffen werden kann. Zudem verhindern sie häufig die Weitergabe der von Forschenden für eine Publikation verwendeten Daten und damit die Reproduzierbarkeit.
9. ***Duplikation:*** Es existieren oft Duplikate von Inhalten, die sich nur im Falle exakter Duplikate leicht durch Software identifizieren lassen.
10. ***Unklarheit über Zugehörigkeit zur Grundgesamtheit:*** Informationen zur Beurteilung, ob ein Fall zur Grundgesamtheit zählt, fehlen in einigen Fällen (z. B. Entstehungsdatum und Entstehungsort eines Dokuments).

---

1 Bei einzelnen Plattformen (z. B. Twitter) gibt es *protected accounts*, die nicht auswertbar sind, sodass nicht immer tatsächlich Daten von allen Nutzenden für Forschungsprojekte verfügbar sind. Zudem gibt es auf manchen Plattformen (z. B. Wikipedia) auch Beiträge, die keinem Nutzenden-Account zuzuordnen sind (z. B. Beiträge von nicht-registrierten Nutzenden).

Beim Typ B (Studien ***mit*** Zustimmung der Teilnehmenden) werden insbesondere folgende ***Probleme und Herausforderungen*** identifiziert:

1. ***Fehlende Stichprobenpläne:*** Aufgrund fehlender Stichprobenpläne (z. B. für Nutzende einer bestimmten Technologie) werden häufig willkürliche oder bewusste Auswahlen oder Schneeballverfahren verwendet. Verzerrungen sind bei solchen Stichproben nur für Merkmale quantifizierbar, die sowohl in der Studie erhoben wurden als auch aus anderen Quellen unverzerrt für die Grundgesamtheit vorliegen.
2. ***Schwierige Trennung des Sampling Error vom Nonresponse Error:*** Bei dieser Art von Stichproben ist der *Sampling Error* zudem nicht sauber vom *Nonresponse Error* zu trennen, da jenseits der freiwillig teilnehmenden Personen oft gar keine Stichprobe existiert. Oft willigt nur ein sehr kleiner Anteil der Grundgesamtheit in eine solche Teilnahme ein.
3. ***Zuordnung von Personen zu Endgeräten:*** Endgeräte können von mehreren Personen verwendet werden (*Clustering*) oder Personen über mehrere Endgeräte verfügen (*Duplication)*. Wird dies nicht erhoben und in Analysen berücksichtigt, kann es zu Verzerrungen kommen.
4. ***Kommerzielle Access-Panels:*** Zum Teil kommen auch kommerzielle Access-Panels für die Datenerhebung oder Rekrutierung zum Einsatz, die ebenfalls nicht auf Stichproben mit bekannten oder abschätzbaren Fehlern basieren.

2.3 Nonresponse/Missing Data Error

Während im Rahmen von Survey-Forschung meist[2] relativ klar zwischen ***Unit Nonresponse*** (eine angefragte Zielperson nimmt nicht an der Befragung teil) und ***Item Nonresponse*** (die/der Proband:in beantwortet eine Frage nicht) unterschieden werden kann, fällt diese Unterscheidung bei unstrukturierten Daten oft schwerer. Die Ursachen für fehlende Daten bei unstrukturierten Daten sind breit gestreut, was anhand der folgenden Beispiele illustriert werden soll:

1. ***Nachträgliche Löschungen*** von nutzergenerierten Inhalten können für fehlende Daten verantwortlich sein, ebenso wie verborgene Inhalte oder gesperrte Nutzende.
2. ***Privacy-Bedenken*** können gerade bei Studien, die die Einwilligung der Proband:innen erfordern, aber auch bei der Social-Media-Nutzung generell, zu systematischer ***Unit Nonresponse*** führen.
3. ***Technische Probleme*** können sowohl die Ursache von fehlenden Aufzeichnungen als auch für *Dropouts* sein.
4. Bei Textdaten kann die ***Veränderung der Zeichencodierung*** zu Problemen führen.
5. Besonders schwerwiegend sind ***systematische Ausfälle***, die mit Merkmalen korrelieren, die eigentlich gemessen werden sollen (z. B. bei deviantem Verhalten). Proband:innen können sich bei speziellen Verhaltensweisen für das Abschalten des Trackings entscheiden und damit die Datenerfassung temporär unterbrechen.
6. Oft ist die ***Ursache der fehlenden Daten unklar***. Bei der Datenerhebung über Smartphones oder Wearables kann möglicherweise nicht unterschieden werden, ob das Gerät vergessen wurde und evtl. Geräusche eines anderen Umfelds aufzeichnet, der Proband bzw. die Probandin schläft und deswegen das Gerät nichts aufgezeichnet oder ein technisches Problem vorliegt.

Zudem ist ***Nonresponse*** bei unstrukturierten Daten auch häufig ***konfundiert mit Undercoverage***. So könnte bei der Analyse von Twitter-Daten in einem festgelegten Zeitraum je nach Definition einer Untersuchungseinheit (Tweet vs. Individuum) die Kategorisierung unterschiedlich ausfallen: Wenn das Individuum als Untersuchungseinheit gewählt wird, kann *Undercoverage* entweder aus einem fehlenden Twitter-Account oder *Nonresponse* im Sinne von ausbleibenden Tweets in dem Zeitraum resultieren. Ist die Untersuchungseinheit ein Tweet, würden alle ausbleibenden Tweets als *Undercoverage* gewertet (Amaya et al., 2020).

2 Die Grauzone bei der Unterscheidung zwischen *Unit* und *Item Nonresponse* in der Survey-Forschung inkludiert etwa Abbrüche bei Online-Befragungen.

2.4 Empfehlungen

Aufgrund der dynamischen Entwicklung und der Unterschiedlichkeit der Datentypen, gestaltet sich die Entwicklung von Standards zum Umgang mit einem Großteil der im Bereich Datengenerierung genannten Probleme schwierig. Die Professionalisierung läuft in den unterschiedlichen Disziplinen parallel oder zeitversetzt. Im Rahmen des oben erwähnten Workshops wurden die folgenden ***Empfehlungen*** abgeleitet:

1. Bei der Verwendung von Plattformdaten sollten Verfahren/Software zur ***Erkennung von Bots*** eingesetzt werden. Es gibt dringenden Bedarf an weiterer Forschung zur Qualität existierender Verfahren (z. B. Rauchfleisch & Kaiser, 2020) und an einer Verbesserung der Erkennung.
2. Die ***Limitationen*** von Studien mit unstrukturierten Daten, insbesondere ***in Bezug auf mögliche Selektivitäten***, sollten in Publikationen ausführlich diskutiert werden.
3. Stehen sehr große Stichproben oder die gesamte Grundgesamtheit für Analysen zur Verfügung, so empfiehlt sich besonders die Interpretation und das Berichten von ***Effektstärkemaßen*** (und ihrer Konfidenzintervalle) zusätzlich zur bzw. anstelle der berichteten statistischen Signifikanz der Ergebnisse.
4. ***Poweranalysen*** zur Bestimmung der Stichprobengröße sind insbesondere dann empfehlenswert, wenn die Kosten pro Untersuchungseinheit hoch sind.
5. Bei der Analyse von Dokumenten sollte eine ***Prüfung auf Duplikate*** stattfinden. Ein ***Fuzzy String Matching*** erlaubt auch die Aufdeckung von nicht exakt identischen Dokumenten (Leskovec et al., 2020).
6. Ein vielversprechender Ansatz, um ***Replizierbarkeit*** von Ergebnissen auf Grundlage schutzwürdiger Daten trotz Datenschutzanforderungen zu ermöglichen, ist der ***Differential-Privacy-Ansatz*** (Dwork et al., 2006; Dwork & Roth, 2014). Dieser wurde z. B. bereits von Evans & King (2022) auf Facebook-URL-Datensatz angewendet (Evans & King, 2022; siehe auch die Möglichkeit von *Remote Execution Solutions* von van Atteveldt et al., 2021).
7. Es gibt Studien zu mobilen Endgeräten wie Smartphones oder Wearables, die ein Mitwirken der Teilnehmenden erfordern. Wenn es aus Gründen der Datenverfügbarkeit, des Datenschutzes oder forschungsökonomischen Gründen nicht möglich ist, mit Zufallsstichproben zu operieren, wird solchen Studien empfohlen, Maßnahmen zu ergreifen, um die ***Verallgemeinerbarkeit der Forschungsergebnisse*** zu erhöhen. Folgende Maßnahmen sind denkbar:
   a. Die Rekrutierung aus und ***Verknüpfung mit existierenden Studien*** (z. B. Längsschnittbefragungen; vgl. Kreuter et al., 2020). Dies ermöglicht die umfassende Untersuchung der Selektivität. Bei Vorhandensein entsprechender Fragen kann außerdem zwischen ***Noncoverage*** (Wer verfügt nicht über die entsprechenden Endgeräte?) und ***Nonresponse*** (Wer nimmt trotz vorhandenen Endgeräts nicht teil?) unterschieden werden (vgl. Keusch et al., 2020; Keusch et al., 2022). Zudem kann die Befragung dabei helfen, Fragen von gemeinschaftlich genutzten Endgeräten und mehreren Endgeräten pro teilnehmende Person aufzuklären.
   b. Bei verdeckten Populationen erlaubt etwa die Anwendung eines ***Respondent Driven Sampling*** (Heckathorn, 1997) die Schätzung von Inklusionswahrscheinlichkeiten.
   c. Die ***Erhebung*** möglichst aussagekräftiger, mit Zielvariablen der Forschung ***korrelierter Informationen*** (nicht ausschließlich demographische Angaben) über die Teilnehmenden, die auch für die Grundgesamtheit bekannt sind (z. B. aus amtlicher Statistik), ermöglicht eine ***effektive Gewichtung***.
8. Technische Ursachen von fehlenden Werten können beispielsweise durch ***hochfrequentes Monitoring*** eingeschränkt werden. Die Validierung von fehlenden Werten kann durch ***Vergleich unterschiedlicher Sensordaten*** erfolgen.
9. ***Coverage Bias und Nonresponse Bias*** sollten grundsätzlich (nicht nur bei unstrukturierten Daten) ***gemeinsam evaluiert*** werden (siehe Eckman & Kreuter, 2017).
10. Die ***Ursachen für die Fehlerquellen*** sollten ***benannt und kategorisiert***, die Kategorisierung transparent dargelegt und die daraus resultierenden Fehler quantifiziert und mit anderen Datenquellen abgeglichen werden (Amaya et al., 2020).

3 Datenaufbereitung

In der klassischen Survey-Forschung werden explizit Instrumente zur Erfassung von Konstrukten entwickelt. Häufig handelt es sich dabei um Items bzw. Fragebögen (siehe ***Design Fragebogen*** in Abbildung 2), die in Bezug auf das zu messende Konstrukt theoriegeleitet formuliert bzw. selektiert werden. Die Forschung mit unstrukturierten Daten greift dagegen vor allem auf vorliegende Daten zurück, die nicht primär für Forschungszwecke generiert wurden. Dies stellt besondere Herausforderungen an die Beurteilung der Qualität der Messungen, die im vorliegenden Kapitel thematisiert werden sollen.

3.1 Spezifikationsfehler und Validität

Ein ***Spezifikationsfehler*** liegt dann vor, wenn das Konstrukt, auf das sich die Forschungsfrage bezieht, nicht genau dem durch die Daten repräsentierten Konstrukt entspricht (Amaya et al., 2020). Dieser Teil des TEF bezieht sich somit auf die Konstruktvalidität. Die ***Konstruktvalidität*** ist dann gegeben, wenn die Schlüsse, die man aufgrund der Daten auf das zugrundeliegende Konstrukt zieht, adäquat und angemessen sind (Eid & Schmidt, 2014; Messick, 1989, 1995). Die Sicherstellung der Konstruktvalidität beginnt in der sozial-, verhaltens- und wirtschaftswissenschaftlichen Forschung üblicherweise schon bei der Test- und Fragebogenkonstruktion. So werden typischerweise zunächst das Konstrukt theoretisch definiert und dann darauf aufbauend Items aufgrund theoretischer Überlegungen formuliert bzw. selektiert, die den Konstruktbereich valide abdecken sollen (Eid & Schmidt, 2014). Im Rahmen umfangreicher Validierungsuntersuchungen wird dann empirisch untersucht, ob die Daten, die anhand des konstruierten Erfassungsinstruments erhoben werden, theoretischen Erwartungen folgen. Aufgrund des primären Rückgriffs auf bereits vorhandene Daten unterscheidet sich die empirische Forschung, die sich unstrukturierter Daten bedient, häufig von diesem prototypischen Vorgehen sozial-, verhaltens- und wirtschaftswissenschaftlicher Forschung:

1. ***Konstruktbezogene Datenauswahl und Konstruktanpassung:*** Geht man von einem a priori definierten Konstrukt aus, das beforscht werden soll, so stellt sich das Problem, unstrukturierte Daten so auszuwählen und weiterzuverarbeiten, dass sie dem a priori definierten Konstrukt möglichst gut entsprechen. Je nach Datenlage kann dies besser oder schlechter gelingen und zu einer Anpassung des ursprünglich fokussierten Konstrukts führen (Sen et al., 2021).
2. ***Bestimmung des Konstrukts durch die Daten:*** Forschung mittels unstrukturierter Daten kann aber auch so angelegt sein, dass die Forschung nicht mit einem a priori definierten Konstrukt startet, sondern die verfügbaren Daten explorativ dahingehend untersucht werden, welche interessanten Konzepte und Konstrukte aufgrund der verfügbaren Daten beforscht werden könnten (Sen et al., 2021). Die Forschungsfragen und die interessierenden Konstrukte werden somit aus den Daten erschlossen und von den verfügbaren Daten bestimmt.
3. ***Definition neuer Konstrukte:*** Die Arbeit mit unstrukturierten Daten könnte auch zur Einführung ganz neuer Konstrukte in die Wissenschaftsgemeinschaft führen, da neue Erfassungsmethoden auch zur Definition und Etablierung neuer Konstrukte führen können (z. B. analog zur Unterscheidung in explizite und implizite Einstellungen in der Sozialpsychologie durch die Berücksichtigung von Reaktionszeiten).

Diese Verwendung unstrukturierter Daten implizieren ***spezifische Probleme und Herausforderungen*** an die Sicherstellung der Konstruktvalidität und der gewählten Validierungsstrategien:

1. ***Fehlende Validitätsstudien:*** Die Untersuchung der Konstruktvalidität eines Fragebogens oder anderen Messinstruments findet idealerweise im Rahmen eines eigenständigen Forschungsprogramms statt, im Rahmen dessen theoriegeleitet Hypothesen in Bezug auf das Verhalten des Messinstruments geprüft werden (Eid & Schmidt, 2014). Solche umfangreicheren Validitätsstudien fehlen häufig noch für unstrukturierte Daten. Die Validität spezifischer Schlüsse kann daher gefährdet sein.
2. ***Fehlender Goldstandard bei der konvergenten Validierung:*** In Bezug auf die ***konvergente Validität*** kann untersucht werden, ob die anhand der unstrukturierten Daten gewonnenen Konstruktmaße in

theoretisch erwartbarer Weise mit anderen Konstruktmaßen zusammenhängen, die anhand anderer Erfassungsmethoden und Datentypen gewonnen wurden. So können zur Validierung von Schlüssen, die anhand unstrukturierter Daten gewonnen werden, z. B. auf Survey-Daten, Interviewdaten oder ethnografischen Daten zurückgegriffen werden (Reveilhac et al., 2022; Tufekci, 2014). Allerdings wird nicht in allen Fällen ein Goldstandardmaß eines Konstrukts existieren, mit dem die unstrukturierten Daten in Beziehung gesetzt werden können. Auch will man mit der Nutzung unstrukturierter Daten häufig Probleme anderer Erfassungsmethoden – wie z. B. dem Selbstbericht – umgehen, um zu valideren Schlüssen zu gelangen. Schließlich stellt sich auch die Frage, ob die mit verschiedenen Methoden erfassten Konstrukte wirklich dasselbe Konstrukt oder unterschiedliche Konstrukte darstellen.

3. ***Sicherung der Inhaltsvalidität (Kontentvalidität):*** Greift man auf vorgefundene Text-, Bild-, Video- und andere Daten zurück und schließt auf allgemeinere Sachverhalte (z. B. die Einstellung der untersuchten Personen), so stellt sich die Frage, wie repräsentativ die Daten für das zu erfassende Konstrukt (z. B. die Einstellung) sind. Dies kann typischerweise nur anhand zusätzlicher Daten und Informationsquellen geprüft werden.
4. ***Analyse der diskriminanten und inkrementellen Validität:*** Passt man die untersuchten Konzepte und Konstrukte an die vorhandenen Daten an bzw. definiert man datengeleitet neue Konstrukte, so stellt sich die Frage, wie sich diese zu etablierten Konstrukten verhalten. Es gilt zu fragen, ob sie hinreichend distinkt zu etablierten Konstrukten sind (diskriminante Validität) und ob sie über etablierte Konstrukte hinausgehend zur besseren Prognose und Erklärung von Phänomenen beitragen (inkrementelle Validität). Auch die Analyse dieser Facetten der Validität erfordert häufig den Einbezug weiterer Daten.

3.2 Messfehler und inhaltliche Fehler

Die Genauigkeit eines Messinstruments wird in den Sozialwissenschaften durch die Reliabilität erfasst. Die Reliabilität ist definiert als der Anteil der Varianz der wahren Werte an der Gesamtvarianz der beobachteten Werte, die sich aus der Varianz der wahren Werte und unsystematischer Varianz zusammensetzt. Sie bringt zum Ausdruck, in welchem Maß sich beobachtete Unterschiede zwischen Personen auf wahre (messfehlerfreie) Unterschiede zurückführen lassen (Eid & Schmidt, 2014; Schnell et al., 2011). Es gibt unterschiedliche Verfahren zur Bestimmung der Reliabilität, die entweder auf die Konsistenz der Messungen innerhalb eines Messinstruments (z. B. Split-half-Methode, Interne Konsistenz, Paralleltest-Methode) oder die Konsistenz der Messungen mit demselben Messinstrument über die Zeit fokussieren (z. B. Test-Retest-Methode). Eine hohe Reliabilität gibt jeweils an, dass die Ergebnisse bei einer wiederholten Messung eines Merkmals hoch miteinander zusammenhängen (d.h. die Messungen korrelieren hoch miteinander) und somit die Messungen nur zu einem geringen Anteil durch unsystematische Varianz überlagert sind.

Eine mangelhafte Reliabilität der Messinstrumente hat Konsequenzen für die weiteren Analysen, da sie zu verzerrten Schätzungen von Zusammenhängen zwischen Merkmalen führen kann. Es können im Prinzip drei Strategien des Umgangs mit Messfehlern in den Sozialwissenschaften unterschieden werden. Erstens wird versucht, bei einer mangelnden Reliabilität die Messinstrumente zu optimieren (z. B. durch den Ausschluss und/oder die Entwicklung neuer Items). Zweitens kann der Messfehler in den weiteren Analysen bei der Interpretation der Ergebnisse berücksichtigt werden (z. B. im Rahmen von Sensitivitätsanalysen). Drittens gibt es Methoden, die explizit versuchen, für den Einfluss messfehlerbehafteter Messungen bei der Schätzung von Zusammenhängen zu korrigieren (z. B. Strukturgleichungsmodelle).

Der Umgang mit potentiell fehlerbehafteten Daten spielt auch bei der Arbeit mit unstrukturierten Daten eine große Rolle. Hierbei ergeben sich verschiedene ***Probleme und Herausforderungen:***

1. ***Anfallende Daten:*** Unstrukturierte Daten liegen häufig in Form von anfallenden Daten (*found data*) vor, für die bis zu einem gewissen Grad unklar bleibt, unter welchen Bedingungen sie entstanden sind. So ist beispielsweise oft nicht genau zu bestimmen, ob verschiedene digitale Spuren von derselben Person stammen oder welche konkreten technischen Einstellungen von einer Plattform vorgenommen wurden. Dies erschwert die Replikation von Messungen sowie die Spezifikation von Messmodellen, die zur Abschätzung der Reliabilität im traditionellen Ansatz benötigt werden.
2. ***Individuelle Erhebungsgeräte:*** Studien mit unstrukturierten Daten setzen häufig technische Messgeräte (z. B. Sensoren oder Tracker) ein oder greifen auf verfügbare Daten zurück, die von den

individuellen Einstellungen (z. B. des Smartphones oder einer Plattform) abhängen können. Es stellt sich somit die Frage, inwiefern Strategien der Sorgfältigkeitsprüfung bestehen, mit denen mögliche Verzerrungen in der Datenerhebung und Datenaufbereitung aufgedeckt und korrigiert werden können.

3. ***Automatisiert erstellte Daten:*** Es besteht die Möglichkeit, dass Daten automatisiert erzeugt werden (z. B. bei Social Media durch Bot-Accounts oder durch automatisierte Inhalte im Webtracking). Fließen solche Daten neben natürlich erhobenen Daten in die Reliabilitätsanalysen ein, könnten diese verzerrt werden.

3.3 Empfehlungen

Es wird insgesamt festgestellt, dass Fragen zur Qualität der Daten einen großen Stellenwert in der Forschung mit unstrukturierten Daten besitzen und ein erheblicher Teil der Zeit in die Aufbereitung und Kontrolle der Daten investiert wird. Allerdings wird auch beobachtet, dass zwar avancierte statistische Verfahren zur Behandlung der Daten eingesetzt werden, sich aber bisher noch kaum allgemeine Standards für die Beurteilung der Qualität der Messungen etablieren konnten. Dies ist u. a. auch darauf zurückzuführen, dass Forschungsteams mit jeweils unterschiedlichen Apps sowie technischen Geräten arbeiten und dass die Plattformen und Apps ständigen Veränderungen unterworfen sind. Auch fehlen häufig systematische Validierungsstudien. Es werden folgende Ansatzpunkte für zukünftige Forschung zur Validität und Reliabilität von unstrukturierten Daten gesehen:

1. ***Rückgriff auf etablierte Validierungsstrategien:*** Studien zur Prüfung der Validität bei unstrukturierten Daten können auf etablierte Verfahren der Sozial-, Wirtschafts- und Verhaltenswissenschaften zurückgreifen (siehe z. B. Eid & Schmidt, 2014; Krippendorff, 2008; Lamnek & Krell, 2016; Schnell et al., 2011) und diese auf unstrukturierte Daten beziehen:
   a. Die ***konvergente Validität*** kann untersucht werden, indem zusätzlich auf andere Datenquellen wie Survey-Daten, Interviewdaten oder ethnografischen Daten zurückgegriffen wird (Reveilhac et al., 2022; Tufekci, 2014). Werden Konstruktmaße anhand von Algorithmen gewonnen, kann anhand externer Daten oder auch speziell generierter Daten überprüft werden, ob der Algorithmus zu ähnlichen Konstruktmaßen führt.
   b. Im Rahmen der ***Kriteriumsvalidierung*** kann analysiert werden, ob ein anhand unstrukturierter Daten bestimmtes Konstruktmaß ein Kriterium erwartungskonform vorhersagt. Außerdem kann analysiert werden, ob Konstruktmaße, die anhand unstrukturierter Daten gewonnen werden, über andere bisher verfügbare Konstruktmaße hinausgehend einen Beitrag zur Vorhersage eines Kriteriums (z. B. bestimmtes Verhalten oder Erleben) liefern und somit einen Zugewinn darstellen ***(inkrementelle Validität)***.
   c. In Bezug auf die ***Inhaltsvalidität (Kontentvalidität)*** kann untersucht werden, wie repräsentativ die ausgewählte Stichprobe von Daten für das zu untersuchende Konstrukt ist. So kann beispielsweise anhand von Topic-Modelling-Verfahren (z. B. Heyer et al., 2018) unter Rückgriff auf Datenbank-Metadaten untersucht werden, ob die gefundene Verteilung von Wörtern in Texten der Verteilung in anderen Datenbanken entspricht. Auch kann es z. B. sinnvoll sein, zu überprüfen, wie repräsentativ die Lebenssituationen, in denen Text-, Audio- oder Bilddaten erhoben wurden, für das Leben der untersuchten Personen sind. Solche Vergleichsdaten könnten über umfangreichere Panelstudien gewonnen werden, die sich explizit auf menschliches Verhalten in der digitalen Welt beziehen sollten (Tufekci, 2014). Mit Hilfe eines Compliance-Fragebogens kann auch erfragt werden, zu welchen Zeiten bzw. in welchen Situationen Personen das Aufzeichnungsgerät nicht getragen haben. Dadurch können mögliche Verzerrungen in der Aufzeichnung aufgedeckt werden.
   d. Bei der inhaltsanalytischen Auswertung von unstrukturierten Daten kann auf Methoden der ***semantischen Validierung*** zurückgegriffen werden (Krippendorff, 2012), um zu überprüfen, ob die Bedeutungsinhalte der analysierten Texte korrekt dargestellt werden.
   e. Werden unstrukturierte Daten durch qualitative Methoden ausgewertet, können ***Validierungsstrategien der qualitativen Forschung*** berücksichtigt werden (z. B. Lamnek & Krell, 2016).

2. ***Fehlende Validitätshinweise:*** Ist es nicht möglich, überzeugende Belege für die Konstruktvalidität anzuführen, so ist die Validität des Schlusses von den Daten auf das zugrundeliegende Konstrukt nicht sichergestellt. In einem solchen Fall
   a. bietet sich es an, diese Schlüsse nicht vorzunehmen und Interpretationen nur in Bezug auf die verfügbaren Daten „datennah" vorzunehmen,
   b. kann es sinnvoll sein, die Fragestellung zu überarbeiten bzw. die Datenquellen zu wechseln (und dies in transparenter Weise zu kommunizieren),
   c. kann es auch notwendig sein, auf die Verwertung der Daten und ihre Publikation ganz zu verzichten.
3. ***Datengesteuerte Konstruktdefinition:*** Wenn die Definition der Konstrukte durch die Daten gesteuert wird, ist es notwendig, eine hohe Transparenz über den Zusammenhang zwischen Konstrukt und Datenpunkten sicherzustellen. So kann anhand von Beispieldaten illustriert werden, in welcher Weise die Daten mit dem interessierenden Konstrukt in Beziehung gesetzt werden. Alternativ wäre an ein stärker ***theoriegestütztes Vorgehen*** zu denken, in dem konkrete Erwartungen über den Zusammenhang zwischen zu erfassendem Konzept und den verfügbaren bzw. erhobenen Daten formuliert werden („Hilfstheorien"; Schnell et al., 2011). Dies würde eine theoriegestützte Validitäts- und Reliabilitätsprüfung erlauben und würde zu einer besseren Integration in die Theorie klassischer Messmodelle führen.
4. ***Reflektion des Einflusses von Messfehlern:*** Die Größe und der mögliche Effekt von Messfehlern sollte kommuniziert und bei der Interpretation der Ergebnisse berücksichtigt werden. Offensichtlich fehlerhafte Datenpunkte sollten ausgeschlossen werden. Die Prüfung der Datenqualität (Messgüte, Messfehlereinflüsse) kann anhand verschiedener Strategien erfolgen, die sich je nach Art der unstrukturierten Daten unterscheiden können.
   a. ***Smartphone-Daten:*** Mit Hilfe von standardisierten Handlungsabläufen kann die ***Korrektheit der aufgezeichneten Daten*** überprüft werden:
      i. Beispielsweise kann ein bestimmtes ***Protokoll*** vorgegeben (App öffnen, Anruf tätigen etc.) und dann mit den aufgezeichneten Daten verglichen werden. Des Weiteren können bestimmte Einstellungen auf Smartphones ***standardisiert*** werden.
      ii. ***Identifikation und sorgfältige Prüfung von Werten, die außerhalb des zu erwartenden Bereichs liegen:*** Dies können beispielsweise Variationen in der Herzrate sein oder sehr schnelle Ortswechsel (mehrere Kilometer in wenigen Sekunden) im GPS.
      iii. ***Messungenauigkeiten*** können teilweise ***protokolliert*** und in den weiteren Analysen berücksichtigt werden. Beispielsweise kann bei der Messung der Geoposition (Longitude und Latitude) auch die Genauigkeit der Messung aufgezeichnet werden.
   b. ***Textdaten:*** Bei der Text- und Inhaltsanalyse lassen sich verschiedene Strategien empfehlen:
      i. ***Pre-Processing von Texten:*** Es können Skripte zur Aufbereitung der Texte und die eingesetzten Algorithmen in Replikationsskripten zur Verfügung gestellt und Plausibilitätsprüfungen durchgeführt werden.
      ii. ***Messfehlerkorrektur bei Inhaltsanalysen:*** Bei Inhaltsanalysen von Social-Media-Daten (z. B. Facebook-Daten) kann die Beobachtungsübereinstimmung der Kodierer:innen zur Bestimmung der Genauigkeit der inhaltlichen Kodierungen verwendet werden (Bachl & Scharkow, 2017). Außerdem können Crowdsourcing-Ansätze genutzt werden, um z. B. Zufallsstichproben von Texten kodieren zu lassen, um die Konvergenz über Kodierer:innen hinweg untersuchen zu können.
      iii. ***Robustheit von Textanalysealgorithmen:*** Die Parameter der eingesetzten Textalgorithmen können variiert werden und somit kann untersucht werden, wie sensitiv die Hauptbefunde gegenüber dieser Variation sind. Beispielsweise kann der Schwellenwert für semantische oder strukturelle Ähnlichkeit in einer Textanalyse variiert werden.
5. Es sollte mehr ***methodische Grundlagenforschung*** zur Entwicklung von Verfahren für die Validitäts- und Reliabilitätsprüfung durchgeführt werden. Das betrifft auch die Publikation von Datenqualitätschecks. Eine Herausforderung wird sein, Verfahren zu entwickeln, die sich auf verschiedene Apps und Plattformen generalisieren lassen.

4 Datenanalyse

Im Gegensatz zur Datengenerierung und Datenaufbereitung beinhaltet die Analyse von unstrukturierten Daten im Prinzip dieselben Arbeitsschritte – und somit auch potentiellen Fehlerquellen – wie die traditionelle Survey-Forschung (siehe Abbildung 2). Ein Großteil der Empfehlungen zur Datenanalyse aus der forschungsmethodischen Literatur zu strukturierten Daten lässt sich somit direkt auf die Arbeit mit unstrukturierten Daten übertragen (z. B. Cohen et al., 2003; Shadish et al., 2002). Im vorliegenden Kapitel wird auf spezifische Herausforderungen bei der Datenanalyse eingegangen, die im Rahmen der von uns durchgeführten Befragung sowie in der daran anschließenden Diskussion im Workshop von den Expert:innen als charakteristisch für unstrukturierte Daten angesehen wurden.

4.1 Record Linkage und Verarbeitungsfehler

Unstrukturierte Daten werden für die weiteren Datenanalysen häufig mit anderen Datenquellen kombiniert (***Record Linkage***). So werden beispielsweise Web-Tracking- und Social-Media-Daten mit klassischen Befragungsdaten verknüpft oder es werden TV-Videoinhalte mit Inhaltsanalysedaten (Link über Sender/Datum/Zeit), Social-Media-Daten mit Webseiteninhalten (Link über URL) und Verhaltensdaten mit Medieninhalten (Link über Tracking-Daten-URLs) miteinander verknüpft (siehe Stier et al., 2020). Für dieses *Linkage* gelten die gleichen Standards wie bei strukturierten Datenformaten (Christen, 2012; Tokle & Bender, 2021). Das Ziel muss sein, mögliche ***Verarbeitungsfehler*** (z. B. keine eindeutige Zuordnung von Personen), die durch das Verknüpfen verschiedener Datenquellen entstehen können, zu minimieren.

Des Weiteren besteht eine besondere Herausforderung für die Datenanalyse darin, dass unstrukturierte Daten überwiegend für ***andere Zwecke als für Forschung*** erhoben werden und daher – um sie für die Forschung nutzbar zu machen – oftmals mit hohem Aufwand in Forschungsdaten transferiert werden müssen. Hierzu muss vor jeder Analyse von unstrukturierten Daten ein Verständnis aufgebaut werden, was diese Daten repräsentieren und wie diese gemessen wurden. Auch ist manchmal die ***genaue Bedeutung***, bzw. der ***Kontext der Erhebung/Generierung der unstrukturierten Daten*** bei manchen Variablen „aus der Distanz" nicht klar. Oftmals liegen beispielsweise keine demographischen Angaben vor und somit können zentrale Bezugsinformationen fehlen (Stier et al., 2020). Außerdem können ***unterschiedliche Geräte*** von verschiedenen Herstellern (z. B. für Aufzeichnungen) zu Unterschieden in den Daten führen, da je nach Gerätetyp unterschiedlich aufgezeichnet wird. Die einzelnen Verarbeitungsschritte – beispielsweise beim *Preprocessing* von Textdaten – müssen daher transparent dokumentiert werden. Bei Textdaten können ***Encoding-Probleme*** oder Unterschiede in den Daten durch unterschiedliche bzw. andere Reihenfolgen von Bereinigungs- bzw. Transformationsschritten entstehen. Dies unterstreicht die Bedeutung einer detaillierten Dokumentation der Programmierschritte und sollte weniger als ein Programmierfehler angesehen werden. Insgesamt wurde festgehalten, dass zwar vereinzelte Best-Practice-Empfehlungen für die Programmierung existieren, das *Preprocessing* und die damit verbundenen Verarbeitungsschritte aber auch von der konkreten Forschungsfrage abhängig sind.

Eine zusätzliche Besonderheit von unstrukturierten Daten ist, dass neben von Nutzenden generierte Daten auch ***systemgenerierte Daten*** erzeugt werden. Es kann daher vorkommen, dass systemgenerierte Daten fälschlicherweise als nutzergenerierte Daten verarbeitet werden.

4.2 Modellierungsfehler

Eine weitere mögliche Fehlerquelle liegt in der Abhängigkeit der Analyseergebnisse von der Wahl der statistischen Modellierung (siehe Abbildung 2). Der ***Modellierungsfehler*** umfasst sowohl die Möglichkeit einer fehlerhaften Anwendung statistischer Modelle als auch die Tatsache, dass Ergebnisse/Schlussfolgerungen sich substantiell ändern können, wenn unterschiedliche statistische Modellierungen gewählt werden.

Bei der ***fehlerhaften Anwendung statistischer Modelle*** kann es sich um die Wahl eines nicht angemessenen Analysemodells handeln (z. B. falsche Methode zur Berechnung der Standardfehler; West et al., 2016), aber auch um Fehler, die unabsichtlich bei der Spezifikation des Analysemodells entstanden sind. Eine Grundvoraussetzung für die Aufdeckung und Korrektur von Modellierungsfehlern ist die Reproduzierbarkeit der Analyseergebnisse (***Computational Reproducibility***; Stodden et al., 2018). Damit ist gemeint, dass anhand des Analysecodes (z. B. Syntax-File der verwendeten Statistik-Software) und der für die Analysen verwendeten Daten, die Ergebnisse reproduziert werden können (Christensen, 2018; Hardwicke et al., 2021).

Zusätzlich stellt sich die Frage nach der ***Robustheit der Analyseergebnisse***. Ergebnisse werden im Allgemeinen als weniger robust angesehen, wenn sie sensitiv gegenüber der Wahl alternativer Analysemodelle sind (Simonsohn et al., 2020; Young & Holsteen, 2017). Beispielsweise können zentrale Analyseergebnisse davon abhängen, welche Kovariaten in das Modell aufgenommen werden und wie die Effekte der berücksichtigten Kovariaten spezifiziert werden (z. B. lineare vs. nicht-lineare Effekte). Fragen der Robustheit der Analyseergebnisse können am besten adressiert werden, wenn sowohl der Analysecode als auch die Daten der Wissenschaftsgemeinschaft zugänglich sind.

4.3 Analytischer Fehler

Abschließend liegt eine zentrale Fehlerquelle in den Schlussfolgerungen, die auf Basis der Datenauswertung vorgenommen werden (***Analytischer Fehler***). Dabei gilt es zu betonen, dass auch bei der Arbeit mit unstrukturierten Daten die Belastbarkeit der Schlussfolgerungen vom Forschungsdesign abhängt (Salganik, 2018). Ein Königsweg für belastbare kausale Schlussfolgerungen stellen sicherlich auch hier randomisierte, längsschnittlich angelegte Experimente oder Feldstudien dar (Shadish et al., 2002). Sollte eine Randomisierung nicht möglich sein, können durch die Berücksichtigung von Kovariaten oder die geschickte Wahl von Untersuchungsdesigns (z. B. natürliche Experimente oder quasi-experimentelle Designs; Angrist & Pischke, 2009), die Effekte möglicher Störvariablen (*Confounder*) kontrolliert werden. Alternativ kann auch von der Beantwortung kausalanalytischer Fragestellungen Abstand genommen werden und der Fokus auf die Beschreibung von Zusammenhängen gelegt werden.

Ein besonderer Aspekt der Analyse von unstrukturierten Daten ist die *Testfairness* der zur Auswertung eingesetzten statistischen Algorithmen. Mit der ***Testfairness*** ist die systematische Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen gemeint. Diese Benachteiligungen können beispielsweise durch die Verwendung von verzerrten Trainingsdaten (*Sample Bias*) für die Klassifizierungsalgorithmen entstehen (Rodolfa et al., 2021). In der Informatik gewinnt die Fairnessüberprüfung von flexiblen maschinellen Lernverfahren zunehmend an Bedeutung (z. B. im Bereich *Algorithmic Biases/Fairness* in *Artificial Intelligence*). In dem von uns durchgeführten Expert:innen-Workshop bestand Konsens, dass die Überprüfung der Modellfairness auch in sozialwissenschaftlichen Anwendungen eine große Rolle spielen sollte, insbesondere wenn es darum geht, die Verfahren in der Praxis einzusetzen.

4.4 Empfehlungen

Die Nutzbarmachung von unstrukturierten Daten für die Analysen ist mit einem hohen Aufwand verbunden, der oftmals für Dritte nicht wirklich sichtbar ist. Allgemein kommt dem ***Preprocessing*** der Daten eine besondere Rolle für die nachfolgenden statistischen Analysen zu. Folgende Empfehlungen werden aus den Ergebnissen unseres Workshops abgeleitet:

1. ***Dokumentation der Verarbeitungsschritte:*** Bei der Transformation der unstrukturierten Daten in Forschungsdaten sollte jeder Vorverarbeitungsschritt dokumentiert werden und auf mögliche verzerrende Effekte untersucht werden. Manchmal ist das aber schwierig bis unmöglich, da *Preprocessing* auf einer Plattform oder in einer App den Forschenden keine Einblicke in die Prozesse gewährt, da dies proprietäre Vorgänge sind. Softwareversionen können sich auch verändern, die dann ggf. abweichende Ergebnisse produzieren. Derartige Probleme können durch Lösungen für Versions- bzw. Dependency-Management adressiert werden (z. B. „packrat“ in R).
2. ***Programmierung:*** Es kann auch zu fehlerhaften Daten durch einfache Programmierfehler kommen, die dann mangels Plausibilitätschecks nicht gefunden werden. Sinnvoll wäre es, Strategien zur

Reduzierung von Fehlern aus der professionellen Programmierung zu übernehmen. Hier wären z. B. *Pair Programming*, Peer Review von Code und natürlich Replikation, Transparenz und Metaanalysen zu nennen. Eine effiziente Datenaufbereitung erfordert aber oft z. B. tiefergreifende Kenntnisse in der objektorientierten Programmierung. Kaum ein:e Sozialwissenschaftler:in hat diese Kenntnisse und umfassend ausgebildete Software-Ingenieur:innen sind für die Forschung schwer zu rekrutieren.

3. ***System- vs. nutzergenerierte Daten:*** Um systemgenerierte von nutzergenerierten Daten zu differenzieren, können nachfolgende Herangehensweisen hilfreich sein:

   a. ***Triangulation:*** Durch anschließende qualitative Interviews kann in Erfahrung gebracht werden, inwieweit den Untersuchungsteilnehmenden ihr Verhalten in bestimmten Situationen bewusst ist (z. B. beim Eye-Tracking: Ist es einer Person überhaupt bewusst, wenn ihre Augen längere Zeit auf einem Gegenstand verweilen und wenn ja, warum verweilen sie dort?). Hieraus lassen sich zusätzliche Informationen gewinnen, die zu einer Validierung der Daten beitragen können.

   b. ***Methodenintegration:*** In textanalytischen Verfahren gibt es die Möglichkeit der Expertenvalidierung. Themenspezifische unstrukturierte Daten (z. B. zu einer politischen Situation in einer bestimmten Region) können durch fachliche Expertise untermauert werden. Eine solche Methodenintegration wäre insbesondere aus Sicht der qualitativ Forschenden erstrebenswert.

   c. ***Methodenkombination:*** Hierbei werden die Daten mit zusätzlichem Wissen aus klassischen Textanalysen angereichert. In den Kommunikationswissenschaften werden beispielsweise u. a. Diskursanalysen zu strittigen Themen (wie z. B. zu Internetregulierung) durchgeführt. Hierbei können große Datenmengen mit Hilfe von computergestützten Verfahren registriert werden. Trotzdem ist es wichtig, dass man anschließend kleine Strichproben herausnimmt und versucht, die groben Muster zu verstehen.

4. ***Transparenz von Analysen:*** Sowohl im Sinne der Reproduzierbarkeit als auch der Robustheit der Analysen sollten der Analysecode und die Daten der Wissenschaftsgemeinschaft zur Verfügung gestellt werden. Allerdings wird dies für manche Datensätze aus datenschutzrechtlichen Gründen nicht möglich sein. In diesem Fall gilt es, über Alternativen nachzudenken (z. B. synthetische Daten oder einen eingeschränkten Datenzugriff; van Atteveldt et al., 2021).

Ausblick: Offene Fragen und Herausforderungen bei der Forschung mit unstrukturierten Daten

Die wissenschaftliche Analyse neuer, unstrukturierter Datentypen führt auch zu weiteren Fragen jenseits des spezifischen Forschungsprozesses. Welche neuen Herausforderungen stellen sich in Bezug auf den Zugang zu Daten? Wie transparent ist der Selektionsprozess? Wie sollten Governance-Strukturen gestaltet sein und welche Ressourcen benötigen Wissenschaftler:innen für die Forschung mit unstrukturierten Daten?

5.1. Datenzugang

Unstrukturierte Daten beinhalten ein breites Spektrum an verschiedenen Datentypen. Dies reicht von Zeitungsseiten über Bilder und Videos bis hin zu Sensordaten. Außerdem stammen sie häufig von sehr verschiedenen, oft wissenschaftsfremden und für Profit agierenden, internationalen Unternehmen. Diese haben häufig kein Interesse, überhaupt das Vorhandensein der Daten für die Wissenschaft öffentlich zu machen. Hinzu kommt, dass sich sowohl die Datenstruktur als auch der Zugang zu den Daten technisch und juristisch kontinuierlich ändert. Dies stellt Forschende vor große technische, rechtliche und prozedurale Herausforderungen (Breuer et al., 2020). Die zeitliche und gegenstandsbezogene Varianz erschwert es zunächst, klare und längerfristig gültige Empfehlungen oder gar Lehrbücher zu erstellen.

Allerdings ist die Problemlage über die verschiedenen sozial-, verhaltens- und wirtschaftswissenschaftlichen Disziplinen hinweg sehr ähnlich. Gepaart mit der schnellen Veränderung der Datenstrukturen und des -zugangs führt dies dazu, dass ***Möglichkeiten zur interdisziplinären Vernetzung*** und des institutionalisierten aktuellen Austauschs zu konkreten Forschungsvorhaben die beste Antwort auf das sich rasch entwickelnde Feld der Analyse unstrukturierter Daten ist. Einzelne Disziplinen wie die Publizistik und Kommunikationswissenschaften, in denen gerade unter den Forschenden, die mit Social Media arbeiten, ein Austausch stattfindet, sind hier Vorreiter. Darüber hinaus sind die Kommunikationswissenschaftler:innen auch in unterschiedlichen Initiativen innerhalb der Fachgesellschaft eingebunden, in denen Empfehlungen zu diesen Themen erarbeitet werden. Ein solcher Austausch sollte für alle im RatSWD vertretenen Fachgesellschaften eröffnet und verstetigt werden, um auch für Forschende der anderen Disziplinen die Arbeit mit unstrukturierten Daten zu erleichtern. Darüber hinaus wären ***methoden- bzw. datenbasierte interdisziplinäre Konferenzen*** für einen Austausch zu Themen wie Datenteilung und Datennachnutzung von unstrukturierten Daten sinnvoll.

Eine weitere Herausforderung stellt die rechtliche Absicherung bei der Forschung mit Daten von privaten Plattformen dar. Universitäten sollten ***Hilfspersonal für das Datenmanagement*** bereitstellen und bei ***urheberrechtlichen und datenschutzrechtlichen Fragen*** helfen.

Zuletzt muss das ***Anreizsystem für Wissenschaftler:innen*** geändert werden, um das Teilen und Dokumentieren unstrukturierter Daten, welches häufig mit einem hohen Aufwand verbunden ist, attraktiver zu machen. Während dies in Fächern wie der Psychologie schon weitverbreitet ist, ist diese Tradition in anderen Disziplinen noch nicht verstetigt. In den Fachgesellschaften und in der Ausbildungs- und Qualifikationsphase wird dem Thema noch zu wenig Aufmerksamkeit geschenkt. Es werden hier allerdings durch die Anforderungen von Forschungsförderern und Publikationsorganen neue Anreize gesetzt. Auch Zeitschriften (u. a. in den Politikwissenschaften) wandeln sich zunehmend und schaffen beispielsweise mit Research-Note-Formaten Anreize, neue Datensätze vorzustellen und deren Analysepotential kurz zu beschreiben. Ein solches Format, welches die Zitierbarkeit und Reputation für die Erstellung von Datensätzen erhöht, wäre auch für Datensätze mit unstrukturierten Daten wünschenswert. Allerdings bräuchte es gerade bei unstrukturierten Daten Anlaufstellen für die Absicherung hinsichtlich der Frage, ob die Daten veröffentlicht werden dürfen.

5.2. Transparenz

Da unstrukturierte Daten häufig von wissenschaftsfernen Digitalunternehmen generiert und zur Verfügung gestellt werden, sind die üblichen Transparenzkriterien schwer einzuhalten. Forschende erhalten Zugang über unbekannte APIs, wissen aber oft nicht, wie die Grundgesamtheit aussieht und welche Selektionskriterien verwendet wurden, um die ihnen zur Verfügung gestellten Daten zu ermitteln. Dokumentationen sind häufig unzureichend und Metadaten werden kaum oder nicht mit ausreichender Genauigkeit zur Verfügung gestellt. Hinzu kommt die Problematik der Schnelllebigkeit. Plattformen verändern ihre Selektionsalgorithmen oder die Datenstruktur selbst und dokumentieren dies oft nicht oder nur unzureichend. Dies macht die Daten sehr flüchtig und stellt die Replizierbarkeit von Studien in Frage.

Die Natur von unstrukturierten Daten selbst führt also dazu, dass Forschende bestimmte Gütekriterien, die sich in der Analyse mit selbstgenerierten Daten bewährt haben, nur mit großem Aufwand oder gar nicht erfüllen können. So stellen vorliegende Templates für Fragenkataloge für die ***Pre-Registrierung*** von Studien mit unstrukturierten Daten häufig eine Hürde dar, weil sie nicht immer übertragbar sind. Gerade bei explorativer bzw. datengetriebener Forschung ist eine Pre-Registierung häufig schwierig. Dennoch sind gerade die vielen Vorverarbeitungsschritte und unterschiedlichen Operationalisierungsmöglichkeiten bei der Forschung mit unstrukturierten Daten ein wichtiges Argument für eine Pre-Registrierung. Dies erlaubt es, den Forschenden sich bereits im Vorfeld Gedanken zu machen, Entscheidungen transparent zu dokumentieren (und ggf. zu revidieren) und im besten Fall bereits konstruktives Feedback von Kolleg:innen zu erhalten. Sollte eine Pre-Registrierung nicht möglich/nicht erfolgt sein, sollten die Arbeitsschritte zumindest retrospektiv dokumentiert und mitveröffentlicht werden, um die Transparenz so weit wie möglich zu erhöhen. Dies ist gerade bei der Arbeit mit unstrukturierten Daten wichtig.

In Hinblick auf die *FAIR-Prinzipien*, denen zufolge Daten auffindbar (*findable*), zugänglich (*accessible*), interoperabel (*interoperable*) und wiederverwendbar (*reusable*) sein sollen, ist die Wiederverwendbarkeit der Daten eine zentrale Frage. Die Forschung mit unstrukturierten Daten ist oft sehr spezifisch und die Daten dürfen häufig nicht vollständig geteilt werden. Dies erhöht noch einmal die Anforderungen an die ***Dokumentation der Datenverarbeitung und die Verfügungstellung der Metadaten***, um anderen Forschenden eine Replikation zumindest mit selbst bei der Plattform angefragten Daten zu ermöglichen. Sowohl die Projektförderung als auch die Strukturfinanzierung von Wissenschaftseinrichtungen sollte diesem Umstand Rechnung tragen.

5.3 Governance

Mit der breiten Nutzung von unstrukturierten Daten stellen sich auch neue Fragen hinsichtlich der Etablierung eines angemessenen institutionellen Rahmens wissenschaftlichen Arbeitens. Angemessene Governance-Ansätze müssen dabei unterschiedliche Ziele und Anforderungen im Blick behalten: Wie werden die schutzwürdigen Interessen der Nutzenden gesichert, die Datenspuren hinterlassen? Wie berücksichtigt man die legitimen Interessen der Organisationen, Unternehmen und staatlicher Akteur:innen, die unstrukturierte Daten erzeugen, speichern und analysieren? Wie sichert man Wisssenschaftler:innen den Zugang zu gesellschaftlich relevanten Beständen unstrukturierter Daten? Und schließlich: Wie verständigt man sich auf Regeln guter wissenschaftlicher Praxis und stellt sicher, dass diese in der Breite eingehalten werden?

Angemessene Governance-Modelle müssen dabei zwei Herausforderungen bewältigen. Zum einen sind in ihrer Zielsetzung sehr unterschiedliche Normen in Einklang zu bringen. Das gilt für ***datenschutzrechtliche Ansprüche***, etwa mit Blick auf die Anonymisierung und Pseudonymisierung von Daten, die eine wissenschaftliche Nutzung aufwändiger machen, bestimmte analytische Zugänge versperren und die gemeinsame Nutzung von Daten erschweren können. Das gilt außerdem für einen angemessenen Ausgleich zwischen den Eigentumsrechten der Unternehmen, in deren Wertschöpfungsnetzwerken unstrukturierte Daten entstehen, und dem grundrechtlich abgesicherten Interesse der Wissenschaft, auf gleichsam ökonomisch wertvolle und gesellschaftlich relevante Datenbestände zuzugreifen.

Zum anderen müssen Governance-Strukturen eine ***Vielzahl unterschiedlicher Akteur:innen*** integrieren. Zu nennen sind zunächst die ***Wissenschaftler:innen*** selbst, die sich in einem dynamisch wachsenden Feld entlang sehr unterschiedlicher Zugänge und Fragestellungen mit unstrukturierten Daten beschäftigen.

Hier beobachten wir eine heterogene wissenschaftliche Praxis, die auf die skizzierten Regulierungsziele unterschiedliche Antworten finden.

Zu berücksichtigen sind weiterhin unterschiedliche ***Wissenschaftsorganisationen***. Darunter zählen Universitäten und Forschungseinrichtungen, die ihren Mitgliedern, bezogen auf Regeln und Ressourcen, einen Rahmen im Umgang mit unstrukturierten Daten setzen. Darüber hinaus sind wissenschaftliche Fachgesellschaften relevante Instanzen in der Aushandlung guter wissenschaftlicher Praxis sowie die Akteur:innen aus der Forschungsförderung, die über Anforderungen etwa in Bezug auf Datenmanagement, die wissenschaftliche Praxis prägen. Von Bedeutung sind zudem wissenschaftliche Fachzeitschriften und andere relevante Publikationsorgane, die über ihre Regelungen zu Manuskripten und der Verfügbarkeit der diesen Veröffentlichungen zugrunde liegenden Daten das Feld prägen.

Mit Blick auf Strukturen und Ressourcen spielen vor allem die ***(wissenschafts-)politischen Akteur:innen*** eine relevante Rolle. Außerdem ist von zentraler Bedeutung, dass unstrukturierte Daten in relevantem Umfang von ***kommerziellen Unternehmen*** und hier speziell von ***global agierenden Plattformen*** generiert werden. Ihre Hoheit über Nutzungsdaten spielt in ihren Geschäftsmodellen eine zentrale Rolle. Dies erschwert die Verfügbarkeit der Datenbestände für wissenschaftliche Zwecke, insbesondere wenn etwa der Zugriff nur über APIs möglich ist, die jederzeit einseitig neu konfiguriert werden können.

Mit ausgeprägten wissenschaftlichen und wissenschaftspolitischen Aktivitäten wie z. B. im Bereich der Nationalen Forschungsdateninfrastruktur (NFDI)[3] wird daran gearbeitet, unstrukturierte Daten systematisch auch für andere Wissenschaftler:innen verfügbar zu machen. Dem steht allerdings aktuell noch eine vergleichsweise geringe Regelungsdichte auf Ebene der wissenschaftlichen Fachgesellschaften gegenüber. Diese beschäftigen sich jenseits der Verfügbarkeit von Daten mit der Frage, mit welchen Methoden valide und reliable Befunde auf den jeweiligen Feldern generiert werden können. So liegen etwa Empfehlungspapiere zur Arbeit mit unstrukturierten Daten in sehr unterschiedlichen Detaillierungsgraden vor. Ein Beispiel sind die Empfehlungen zum Umgang mit Forschungsdaten in der Kommunikationswissenschaft (Peter et al., 2020), die in einer Arbeitsgruppe der Deutschen Gesellschaft für Publizistik und Kommunikationswissenschaft (DGPuK) erarbeitet wurden. Grundsätzlich ist zu empfehlen, den Diskurs in den Fächern zu intensivieren und Formate zu finden, in denen über die jeweiligen Fachspezifika hinaus ein ***gemeinsames Verständnis guter wissenschaftlicher Praxis*** entwickelt wird. Hierbei ist sinnvollerweise auch die Expertise aus Forschungsförderung, Infrastrukturen und Fachpublikationen einzubeziehen.

Bezüglich der Überlegungen über angemessene Governance-Strukturen ist der Bedarf an ***politischer Regulierung hinsichtlich der Kooperation mit kommerziellen Anbietern*** von zentraler Bedeutung. Es ist sowohl problematisch, dass die Unternehmen ihre Daten selbst kuratieren, als auch, dass sie freiwillig entscheiden können, ob sie mit der Wissenschaft kooperieren. Hierdurch entstehen starke Abhängigkeiten, in denen die ökonomischen und strategischen Interessen von Plattformen signifikanten Einfluss darauf nehmen können, welche Fragestellungen erforscht und welche Befunde analysiert werden. In der aktuellen Debatte zur Regulierung von digitalen Angeboten werden zwar Ansätze erkennbar, die den Zugang zu Daten für Wissenschaftler:innen erreichen können, allerdings bleibt abzuwarten, in welchem Umfang diese Ansätze umgesetzt werden und geeignet sind, Zugangshürden für Wissenschaftler:innen zu senken. Hier besteht daher weiterer Bedarf, seitens der Politik den Druck auf Plattformanbieter zu erhöhen. Es ist jedoch schwierig, politische Forderungen an kommerzielle Unternehmen zu stellen, solange die öffentlichen Stellen selbst keine Vorreiterrolle dabei spielen, Wissenschaftler:innen einfachen Zugang zu Daten zu gewähren, etwa im Bereich der amtlichen Statistik. Hier wäre ein kohärentes Handeln im Sinne des möglichst offenen Zugangs zu Daten wünschenswert. Erste Initiativen in dieser Richtung existieren bereits, etwa das European Digital Media Observatory (EDMO)[4].

5.4. Ressourcen

Die Analyse unstrukturierter Daten erfordert aufgrund ihres Volumens und ihrer zeitlichen Dynamik deutlich mehr Ressourcen als der Umgang mit punktuell erhobenen Datensätzen. Es liegt u.a. daran, dass es deutlich weniger und teilweise keine Standardisierungen gibt und Einzelfallprüfungen in größerem Umfang erforderlich sind.

3 https://www.nfdi.de/
4 https://edmo.eu/

Aus Ressourcenperspektive handelt es sich hierbei weniger um ein technisches Problem, da die notwendigen Technologien zur Archivierung und Analyse auch großer Datenmengen grundsätzlich vorhanden sind. Ebenso können bestehende Infrastrukturen mit genutzt werden und sich ggf. mit vergleichsweise geringem finanziellem Aufwand skalieren lassen. Vielmehr liegt die Herausforderung darin, diese Technologien in technischen Plattformen (z. B. Dataverse[5] im Bereich der Politikwissenschaft) so zu konfigurieren, dass sie für Anwender:innen aus der Wissenschaft einen Mehrwert bieten (z. B. Hemphill, 2019). Dabei geht es nicht nur um den ***Aufbau der Plattform***, sondern vielmehr um deren ***Wartung und kontinuierlichen Anpassung an neue technische Anforderungen***, etwa in Bezug auf Verfügbarkeit, Effizienz, Datenschutz und der Verhinderung von Cyberattacken. Darüber hinaus stehen die inhaltlichen Anforderungen, etwa bezogen auf Auffindbarkeit, Datenqualität, Formatierungen oder Metadaten im Vordergrund (siehe Breuer et al., 2021 für Social-Media-Daten).

Die genannten Herausforderungen lassen sich zu einem Teil mit ***Investitionen in die technische Infrastruktur*** bewältigen. Bedeutsamer ist allerdings die personelle Komponente. Bezogen auf den informationstechnischen Umgang mit Daten sind neue Kompetenzprofile notwendig, die sich in etablierten Stellenbeschreibungen innerhalb von Wissenschaftsorganisationen selten niederschlagen. Selbst wenn es gelingt, hier neue Berufsfelder zu definieren und die entsprechenden Ressourcen für den Betrieb sowie die Aus- und Weiterbildung zur Verfügung zu stellen, sind erhebliche Anstrengungen notwendig, einschlägig qualifiziertes Personal für die Wissenschaft zu gewinnen. Im Wettbewerb mit finanzstarken kommerziellen Akteur:innen scheint es angezeigt, den gesellschaftlichen Mehrwert von wissenschaftlicher Arbeit in den Vordergrund zu rücken, um talentierte Bewerber:innen zumindest für befristete Zeiträume für eine Tätigkeit im Bereich des wissenschaftlichen Datenmanagements zu gewinnen.

Neben den Fragen der ***Rekrutierung und Qualifizierung geeigneter Mitarbeiter:innen*** erfordert die zuverlässige Bereitstellung und Dokumentation von unstrukturierten Daten zusätzliche personelle Ressourcen. In diesem Kontext geht es auch um die absolute Höhe von Budgets. Mindestens ebenso bedeutsam ist die Frage der Finanzierungslogik. Hier kollidieren die Ansprüche einer auf Dauer gestellten Infrastruktur, die dazu notwendig ist, unstrukturierte Daten nutzbar zu machen, mit einer projektorientierten Forschungslogik, die zwar Anreize zu systematischem Datenmanagement setzen, aber die nachhaltige Verfügbarkeit von unstrukturierten Daten nicht vollständig sicherstellen kann. Im Sinne dieser Nachhaltigkeit gilt es einerseits, darüber nachzudenken, welche Support-Funktionen dezentral auf Ebene von Universitäten und Instituten zur Verfügung gestellt werden sollten. Andererseits erscheint es sinnvoll, auf die stärkere Zusammenarbeit von Forschungsinfrastrukturanbietern hinzuwirken, die im Umgang mit unstrukturierten Daten bereits über umfangreiche Erfahrungen verfügen.

Eine weitere Möglichkeit der Qualifikation und des Wissensaufbaus besteht in einer verstärkten ***Kooperation zwischen Sozialwissenschaften und Informatik***. Im Idealfall ergänzen sich die beiden Disziplinen und ein Wissenstransfer findet in beide Richtungen statt. Dies kann sehr produktiv sein, da man sich inhaltlich und in der Vorgehensweise ergänzt: Während in den Sozialwissenschaften vor allem beschrieben und erklärt wird, ist das Ziel in der Informatik die datengetriebene Vorhersage (*Prediction Task*). Obwohl es prinzipiell keine Unterschiede hinsichtlich des Forschungsdesigns (einschließlich Methodenauswahl) und der Qualitätsstandards zwischen sozialwissenschaftlichen oder informatischen Studien gibt, ist die Publikationskultur der beiden Disziplinen verschieden. Es ist daher ratsam, bereits im Vorfeld einer gemeinsamen Arbeit abzustimmen, ob es sich um eine sozialwissenschaftliche oder informatische Studie handelt. Im Rahmen einer solchen interdisziplinären Kooperation besteht die Konfliktgefahr, dass einer der beiden Seiten in die sekundäre Dienstleistungsrolle gerät. Die Herausbildung der ***Computational Social Science*** kann als eine Reaktion auf diese Problematik gesehen werden: Man eignet sich lieber die entsprechenden Kompetenzen an, als eine Dienstleistungsrolle zu übernehmen (Kinder-Kurlanda & Weller, 2014).

5 https://dataverse.harvard.edu/

6 Literaturverzeichnis

Amaya, A., Biemer, P. & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology, 8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Bachl, M. & Scharkow, M. (2017). Correcting measurement error in content analysis. *Communication Methods and Measures, 11*(2), 87–104. https://doi.org/10.1080/19312458.2017.1305103

Biemer, P. & Amaya, A. (2020). Total error frameworks for found data. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov & L. E. Lyberg (Hrsg.), *Big data meets survey science* (S. 131–161). Wiley. https://doi.org/10.1002/9781118976357.ch4

Borkenau, P. (2006). Selbstbericht. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologie: Bd. 4. Handbuch der psychologischen Diagnostik* (S. 135–142). Hogrefe.

Bosch, O. & Revilla, M. (2022). When survey science met online tracking: Presenting an error framework for metered data. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 1-29. https://doi.org/10.1111/rssa.12956

Bosse, S., Dahlhaus, L. & Engel, U. (2021). Web data mining: Collecting textual data from web pages using R. In U. Engel, A. Quan-Haase, S. Liu & L. Lyberg (Hrsg.), *Handbook of computational social science: Data science, statistical modelling, and machine learning methods* (S. 46–70). Routledge.

Breuer, J., Bishop, L. & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society, 22*(11), 2058–2080. https://doi.org/10.1177/1461444820924622

Breuer, J., Borschewski, K., Bishop, L., Vávra, M., Štebe, J., Strapcova, K. & Hegedűs, P. (2021). *Archiving social media data: A guide for archivists and researchers.* https://doi.org/10.5281/ZENODO.5041072

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer. https://doi.org/10.1007/978-3-642-31164-2

Christensen, G. (2018). *Manual of best practices in transparent social science research.* UC Berkeley Initiative for Transparency in the Social Sciences. https://github.com/garretchristensen/BestPractices-Manual/blob/master/Manual.pdf (Zugriff am 02.08.2022).

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3. Aufl.). Erlbaum.

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Hrsg.), *Theory of cryptography. TCC 2006. Lecture notes in computer science* (Bd. 3876, S. 265–284). Springer. https://doi.org/10.1007/11681878_14

Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3-4), 211–407. https://doi.org/10.1561/0400000042

Eberendu, A. (2016). Unstructured data: An overview of the data of big data. *International Journal of Computer Trends and Technology (IJCTT), 38*(1), 46–50. https://doi.org/10.14445/22312803/IJCTT-V38P109

Eckman, S. & Kreuter, F. (2017). The undercoverage-nonresponse tradeoff. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker & B. T. West (Hrsg.), *Total survey error in practice* (S. 95–113). Wiley. https://doi.org/10.1002/9781119041702.ch5

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion.* Hogrefe.

Evans, G. & King, G. (2022). Statistically valid inferences from differentially private data releases, with application to the facebook URLs dataset. *Political Analysis, 31*(1), 1–21. https://doi.org/10.1017/pan.2022.1

Fischer, C., Fishman, B. & Schoenebeck, S. Y. (2019). New contexts for professional learning: Analyzing high school science teachers' engagement on Twitter. *AERA Open, 5*(4). https://doi.org/10.1177/2332858419894252

Fritsche, I. & Linneweber, V. (2006). Nonreactive methods in psychological research. In M. Eid & E. Diener (Hrsg.), *Handbook of multimethod measurement in psychology* (S. 189–203). American Psychological Association. https://doi.org/10.1037/11383-014

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Grimmer, J., Roberts, M. & Stewart, B. (2022). *Text as data: A new framework for machine learning and the social sciences.* Princeton University Press.

Groves, R. M. & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hardwicke, T., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M., Peloquin, B., deMayo, B., Long, B., Yoon, E. & Frank, M. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: An observational study. *Royal Society Open Science, 8*(1), 201494. https://doi.org/10.1098/rsos.201494

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44*(2), 174–199. https://doi.org/10.2307/3096941

Hemphill, L. (2019). *Updates on ICPSR's Social Media Archive (SOMAR).* https://doi.org/10.5281/ZENODO.3612676

Heyer, G., Wiedemann, G. & Niekler, A. (2018). Topic-Modelle und ihr Potenzial für die philologische Forschung. In H. Lobin, R. Schneider & A. Witt (Hrsg.), *Digitale Infrastrukturen für die germanistische Forschung* (S. 351–368). De Gruyter. https://doi.org/10.1515/9783110538663-016

Hox, J. (2017). Computational social science methodology, anyone? *Methodology, 13*(Suppl. 1), 3–12. https://doi.org/10.1027/1614-2241/a000127

Hsieh, Y. P. & Murphy, J. (2017). Total Twitter error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker & B. T. West (Hrsg.), *Total survey error in practice* (S. 23–46). Wiley. https://doi.org/10.1002/9781119041702.ch2

Jürgens, P., Stark, B. & Magin, M. (2020). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review, 38*(5), 600–615. https://doi.org/10.1177/0894439319831643

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F. & Trappmann, M. (2020). Coverage error in data collection combining mobile surveys with passive measurement using apps: Data from a German national survey. *Sociological Methods & Research*, 1-38. https://doi.org/10.1177/0049124120914924

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., Trappmann, M. & Eckman, S. (2022). Non-participation in smartphone data collection using research apps. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185*(Suppl. 2), S225-S245. https://doi.org/10.1111/rssa.12827

Kinder-Kurlanda, K. & Weller, K. (2014). „I always feel it must be great to be a hacker!". In F. Menczer, J. Hendler, W. Dutton, M. Strohmaier, E. T. Meyer & C. Cattuto (Hrsg.), *Proceedings of the 2014 ACM conference on web science - WebSci ,14*, Bloomington, Indiana, USA. 6/23/2014 - 6/26/2014 (S. 91–98). ACM Press. https://doi.org/10.1145/2615569.2615685

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S. & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review, 38*(5), 533–549. https://doi.org/10.1177/0894439318816389

Krippendorff, K. (2008). Validity. In W. Donsbach (Hrsg.), *The international encyclopedia of communication.* Wiley.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (3. Aufl.). Sage Publications.

Lamnek, S. & Krell, C. (2016). *Qualitative Sozialforschung* (6. Aufl.). Beltz.

Lazer, D. & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology, 43*(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Leskovec, J., Rajaraman, A. & Ullman, J. D. (2020). *Mining of massive datasets* (3. Aufl.). Cambridge University Press. https://doi.org/10.1017/9781108684163

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research. *Policy & Internet, 5*(3), 328–339. https://doi.org/10.1002/1944-2866.POI336

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (3. Aufl., S. 13–103). Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 400–408. https://doi.org/10.1609/icwsm.v7i1.14401

Nyhuis, D. (2021). Application programming interfaces and web data for social research. In U. Engel, A. Quan-Haase, S. Liu & L. Lyberg (Hrsg.), *Handbook of computational social science: Data science, statistical modelling, and machine learning methods* (S. 33–45). Routledge.

Peter, C., Breuer, J., Masur, P. K., Scharkow, M. & Schwarzenegger, C. (2020). Empfehlungen zum Umgang mit Forschungsdaten in der Kommunikationswissenschaft: AG Forschungsdaten im Auftrag des Vorstands der DGPuK. *Studies in Communication and Media, 9*(4), 599–626. https://doi.org/10.5771/2192-4007-2020-4-599

RatSWD (Rat für Sozial- und Wirtschaftsdaten). (2020). *Datenerhebung mit neuer Informationstechnologie: Empfehlungen zu Datenqualität und -management, Forschungsethik und Datenschutz* (Output Series, 6. Berufungsperiode Nr. 6). Berlin. https://doi.org/10.17620/02671.47

Rauchfleisch, A. & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PLoS one, 15*(10), e0241045. https://doi.org/10.1371/journal.pone.0241045

Reveilhac, M., Steinmetz, S. & Morselli, D. (2022). A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion. *Multimedia Tools and Applications, 81*(7), 10107–10142. https://doi.org/10.1007/s11042-022-12101-0

Rodolfa, K. T., Saleiro, P. & Ghani, R. (2021). Bias and fairness. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter & J. Lane (Hrsg.), *Big data and social science: Data science methods and tools for research and practice* (2. Aufl., S. 281-312). CRC Press.

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age.* Princeton University Press. http://bvbr.bib-bvb.de:8991/exlibris/aleph/a23_1/apache_media/I2PSYRNNQYGGFGRYQPHL6XDRAYAUGJ.pdf

Scharkow, M. (2016). The accuracy of self-reported internet use: A validation study using client log data. *Communication Methods and Measures, 10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Schnell, R., Hill, P. B. & Esser, E. (2011). *Methoden der empirischen Sozialforschung* (9. Aufl.). Oldenbourg.

Sen, I., Flöck, F., Weller, K., Weiß, B. & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly, 85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company.

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behavior, 4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Stier, S., Breuer, J., Siegers, P. & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review, 38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stodden, V., Seiler, J. & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Taleb, I., Serhani, M. A. & Dssouli, R. (2018). Big data quality assessment model for unstructured data. *2018 International Conference on Innovations in Information Technology (IIT)*, 69–74. https://doi.org/10.1109/INNOVATIONS.2018.8605945

Tanwar, M., Duggal, R. & Khatri, S. K. (2015). Unravelling unstructured data: A wealth of information in big data. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 1–6. https://doi.org/10.1109/ICRITO.2015.7359270

Tokle, J. & Bender, S. (2021). Record linkage. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter & J. Lane (Hrsg.), *Big data and social science: Data science methods and tools for research and practice* (2. Aufl., S. 43–65). CRC Press.

Tufekci, Z. (2014). *Big questions for social media big data: Representativeness, validity and other methodological pitfalls: ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*. https://arxiv.org/pdf/1403.7400

van Atteveldt, W., Althaus, S. & Wessler, H. (2021). The trouble with sharing your privates: Pursuing ethical open science and collaborative research across national jurisdictions using sensitive data. *Political Communication, 38*(1-2), 192–198. https://doi.org/10.1080/10584609.2020.1744780

Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Rand McNally.

West, B. T., Sakshaug, J. W. & Aurelien, G. A. S. (2016). How big of a problem is analytic error in secondary analyses of survey data? *PLoS one, 11*(6), e0158120. https://doi.org/10.1371/journal.pone.0158120

Young, C. & Holsteen, K. (2017). Model uncertainty and robustness. *Sociological Methods & Research, 46*(1), 3–40. https://doi.org/10.1177/0049124115610347

Mitwirkende bei der Erstellung

Mitglieder der AG Herausforderungen bei der wissenschaftlichen Erhebung und Nutzung unstrukturierter Daten

Prof. Stefan Bender
Deutsche Bundesbank, RatSWD

Prof. Dr. Michael Eid, *Co-Vorsitz*
Freie Universität Berlin, RatSWD

Prof. Dr. Christiane Gross
Julius-Maximilians-Universität Würzburg, RatSWD

Prof. Dr. Stefan Liebig
Freie Universität Berlin

Prof. Dr. Oliver Lüdtke, *Co-Vorsitz*
Leibniz-Institut für Pädagogik der Naturwissenschaften und Mathematik (IPN), Christian-Albrechts-Universität zu Kiel, RatSWD

Prof. Dr. Laura Seelkopf
Ludwig-Maximilians-Universität München, RatSWD

Prof. Dr. Lars Rinsdorf
Hochschule der Medien Stuttgart

Prof. Dr. Mark Trappmann
Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA), Otto-Friedrich-Universität Bamberg, RatSWD

Konsultation

Dr. Johannes Breuer
GESIS – Leibniz-Institut für Sozialwissenschaften

Prof. Dr. Christian Fischer
Eberhard Karls Universität Tübingen

Dr. Stephanie Geise
Westfälische Wilhelms-Universität Münster

Dr. Theresa Gessler
Universität Zürich

Dr. Fenne große Deters
Universität Potsdam

Dr. Pascal Jürgens
Johannes Gutenberg-Universität Mainz

Prof. Dr. Florian Keusch
Universität Mannheim

Prof. Dr. Wenzel Matiaske
Helmut-Schmidt-Universität Hamburg

Prof. Matthias Mehl, PhD.
University of Arizona

Prof. Dr. Jürgen Pfeffer
Technische Universität München

Julia Rakers
Universität Duisburg-Essen

Christian Strippel
Freie Universität Berlin

Dr. Katrin Weller
GESIS – Leibniz-Institut für Sozialwissenschaften

Geschäftsstelle RatSWD

Marie Eilers

Impressum

Herausgeber:
Rat für Sozial- und Wirtschaftsdaten (RatSWD)
Geschäftsstelle
Am Friedrichshain 22
10407 Berlin
office@ratswd.de
https://www.ratswd.de

Redaktion:
Marie Eilers

Gestaltung/Satz:
Claudia Kreutz

Berlin, Februar 2023

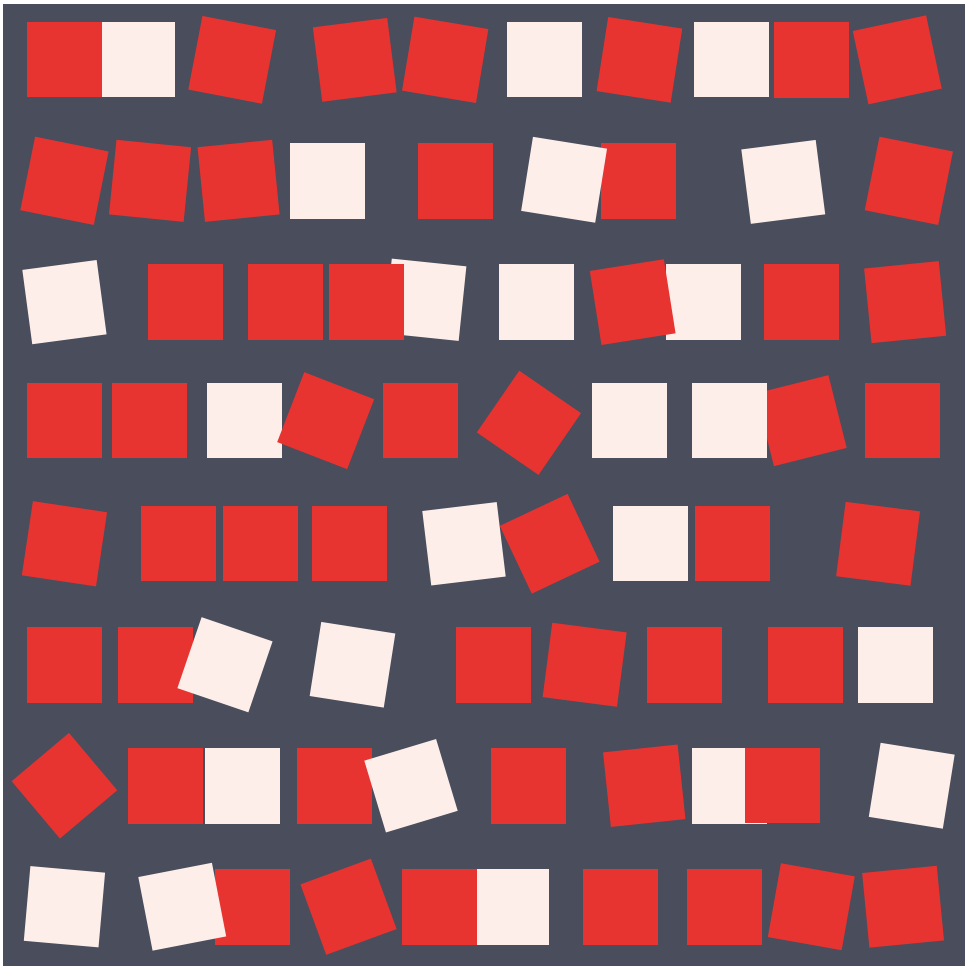
RatSWD Output:
Die RatSWD Output Series dokumentiert die Arbeit des RatSWD in seiner 7. Berufungsperiode (2020–2023). In ihr werden seine Stellungnahmen und Empfehlungen veröffentlicht und auf diesem Weg einer breiten Leserschaft zugänglich gemacht.
Die Geschäftsstelle des RatSWD wird als Teil von KonsortSWD im Rahmen der NFDI durch die Deutsche Forschungsgemeinschaft (DFG) gefördert - Projektnummer: 442494171.

DOI: 10.17620/02671.73

Zitationsvorschlag:
RatSWD (Rat für Sozial- und Wirtschaftsdaten) (2023). *Erhebung und Nutzung unstrukturierter Daten in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Herausforderungen und Empfehlungen.* (RatSWD Output Series, 7. Berufungsperiode Nr. 2). Berlin. https://doi.org/10.17620/02671.73

Der Rat für Sozial- und Wirtschaftsdaten (RatSWD) berät seit 2004 die Bundesregierung und die Regierungen der Länder in Fragen der Forschungsdateninfrastruktur für die empirischen Sozial-, Verhaltens- und Wirtschaftswissenschaften. Im RatSWD arbeiten zehn durch Wahl legitimierte Vertreterinnen und Vertreter der sozial-, verhaltens- und wirtschaftswissenschaftlichen Fachdisziplinen mit zehn Vertreterinnen und Vertretern der wichtigsten Datenproduzenten zusammen.

Der RatSWD ist Teil des Konsortiums für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften (KonsortSWD) in der Nationalen Forschungsdateninfrastruktur (NFDI). Er versteht sich als institutionalisiertes Forum des Dialoges zwischen Wissenschaft und Datenproduzenten und erarbeitet Empfehlungen und Stellungnahmen. Dabei engagiert er sich für eine Infrastruktur, die der Wissenschaft einen breiten, flexiblen und sicheren Datenzugang ermöglicht. Diese Daten werden von staatlichen, wissenschaftsgetragenen und privatwirtschaftlichen Akteuren bereitgestellt. Derzeit hat der RatSWD 42 Forschungsdatenzentren akkreditiert und fördert deren Kooperation.



www.ratswd.de