RatSWD.

Rat für Sozial- und WirtschaftsDaten

Herausforderungen bei unstrukturierten Daten

9. Konferenz für Sozial- und Wirtschaftsdaten (KSWD) 27./28. März 2023

RatSWD Arbeitsgruppe in der 7. Berufungsperiode des RatSWD (2020-2023)

„Herausforderungen bei der wissenschaftlichen Erhebung und Nutzung unstrukturierter Daten“

Michael Eid und Oliver Lüdtke (Vorsitzende der AG)

Qualitätssicherung unstrukturierter Daten

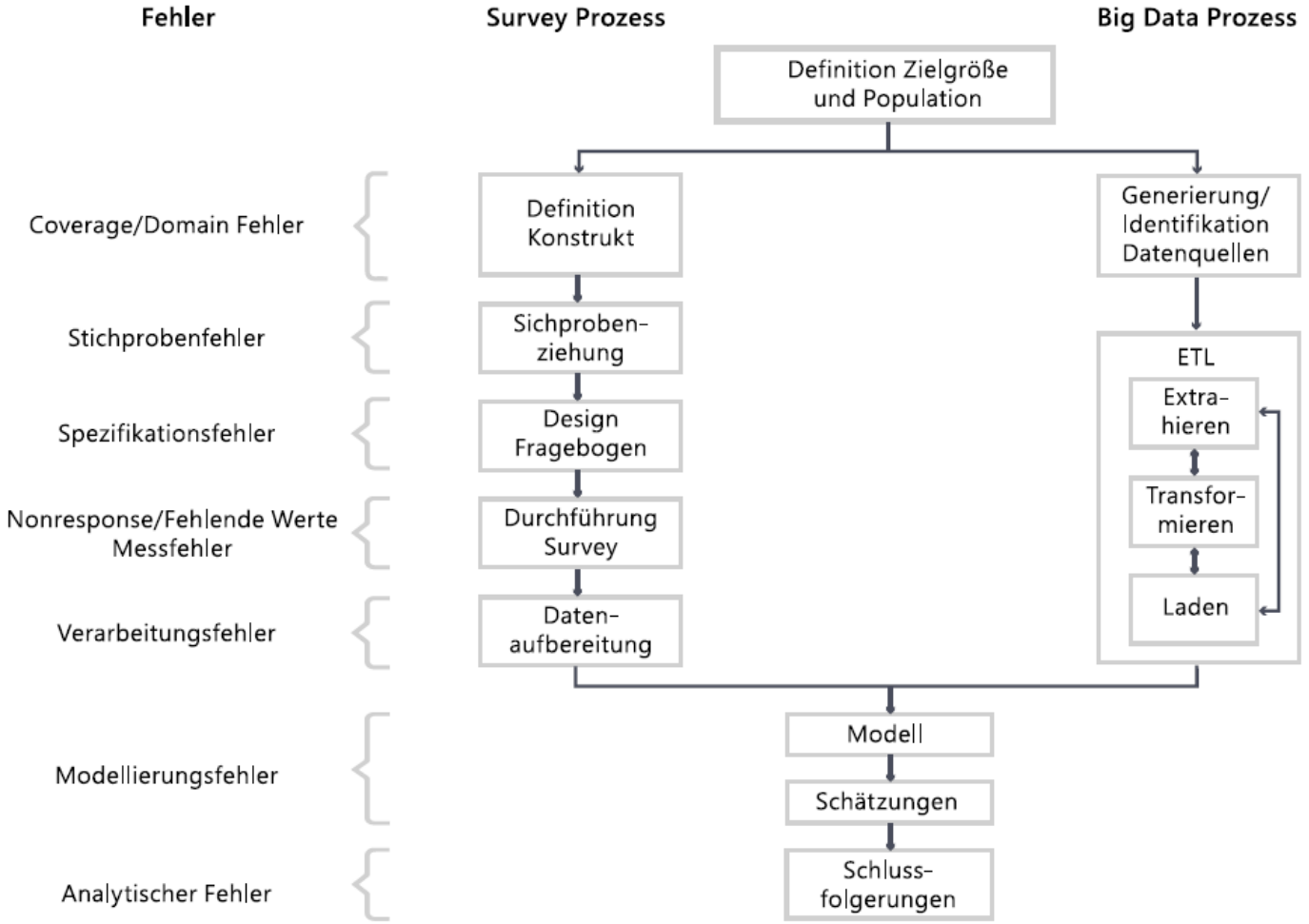
- Zunehmend wird mit Daten geforscht, die durch Nutzung neuer elektronischer Medien entstehen
  - Bilder, Audio- und Videodateien, Textnachrichten, Tweets, Sensordaten etc.
- Diese sog. unstrukturierten Daten sind für die Forschung besonders wertvoll, da sie das menschliche Leben und Verhalten im Alltag abbilden können.
- Evaluation der Qualität dieser Daten besonders herausfordernd, da deren Erhebung häufig nicht von Forschenden konzipiert und durchgeführt wird

Die AG

- Arbeitsgruppe in der **7. Berufungsperiode des RatSWD** (2020-2023)
- 8 Mitglieder
  - Stefan Bender; Michael Eid (Co-Vorsitz); Christiane Gross; Stefan Liebig; Oliver Lüdtke (Co-Vorsitz); Lars Rinsdorf (extern, Hochschule der Medien Stuttgart; DGPuK); Laura Seelkopf; Mark Trappmann
- Ziel: Erarbeitung von **Leitfragen,** Durchführung eines **Workshops** und **Veröffentlichung von Empfehlungen** des RatSWD im Rahmen der RatSWD Output Series

Total Error Framework

Erweiterung des Total Survey Error Framework auf Big Data (Amaya, Biemer & Kinyon, 2020)



Quelle: Amaya et al. (2020)

Fragebogen zur Erhebung und Nutzung unstrukturierter Daten (insgesamt 32 Fragen)

- Allgemeine fachliche Angaben
- Datengenerierung
- Datenaufbereitung
- Datenanalyse
- Offene Abschlussfrage

c) Über Ihre eigene Expertise hinaus: Welche unstrukturierten Datentypen und Fragestellungen sind für Ihre Fachdisziplin üblicherweise typisch?

d) Gibt es in Ihrer Fachdisziplin bereits Richtlinien oder Regeln, die für die wissenschaftliche Nutzung unstrukturierter Daten Anwendung finden? Wenn ja, welche sind das?

2. Angaben zur Datengenerierung

2.1 Im Gegensatz zu Studien, die durch die Forschenden selbst geplant werden, kann die Untersuchungseinheit bei der Erhebung unstrukturierter Daten weniger eindeutig definiert sein. In den folgenden Fragen möchten wir gern erfahren, wie Sie damit umgehen. Nehmen Sie hierzu gern Bezug auf konkrete Beispiele aus Ihren eigenen Studien.

a) Denken Sie an Ihre aktuell durchgeführten Studien mit unstrukturierten Daten: Nach welchen Kriterien definieren Sie die Untersuchungseinheit und die Grundgesamtheit?

b) Welches Stichprobendesign haben Sie gewählt und warum?

c) Welche Rolle spielen unterschiedliche Analyseebenen[1] in Ihrer Arbeit mit unstrukturierten Daten?

d) Gab es hinsichtlich der Verfügbarkeit der Daten irgendwelche Limitationen? Wenn ja, wie gehen Sie damit um und wie beeinflusst dies die Entwicklung des Studiendesigns?

e) Auf einer Skala von 1 (gar nicht relevant) bis 5 (sehr relevant), wie relevant sind diese Fragen für Ihre persönliche Forschungspraxis?

Befragung (September 2021)

- Insgesamt 19 Fragebogen wurden bearbeitet
  - Betriebswirtschaftslehre (1)
  - Bildungsforschung (1)
  - Computational Social Science (1)
  - Kommunikationswissenschaft (5)
  - Politikwissenschaft (2)
  - Psychologie (5)
  - Soziologie (3)
  - Volkswirtschaftslehre (1)

Befragung

- Vielzahl unterschiedlicher Datentypen
  - Social Media (z.B. Twitter, Facebook), Smartphonedaten, Audiodaten
  - klassische Medien (z.B. Zeitungen), Textdokumente (z.B. Parlamentsreden, Betriebsvereinbarungen, Abrechnungen)

Workshop am 13. und 14. Oktober 2021

Tag 1 (13.10.)	
14:00-14:15	Begrüßung
14:15-15:00	Inputrunde der Teilnehmenden
15:00-16:15	**Session 1** Datengenerierung (Gross/Trappmann)
16:15-16:30	Kaffeepause
16:30-17:45	**Session 2** Datenaufbereitung (Eid/Lüdtke)
17:45-18:00	Wrap-up Tag 1

Tag 2 (14.10.)	
14:00-14:10	Begrüßung
14:10-15:25	**Session 3** Datenanalyse (Bender/Liebig)
15:25-15:40	Kaffeepause
15:40-16:55	**Session 4** Offene Fragen zur Verfügbarkeit neuer Datentypen (Seelkopf/Rinsdorf)
16:55-17:40	Abschlussdiskussion

AG-Output

- Bericht „Erhebung und Nutzung unstrukturierter Daten in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Herausforderungen und Empfehlungen“
- Erster Schritt zur Verständigung über Qualitätsstandards für die Arbeit mit unstrukturierten Daten
  - Was sind besondere Herausforderungen, die sich im Vergleich zur Arbeit mit Daten aus traditioneller Survey-Forschung stellen?
  - Welche Ressourcen benötigen Wissenschaftler:innen für die Forschung mit unstrukturierten Daten?

Aus dem Inhalt des Outputs

- Einleitung mit Definition des Begriffs „unstrukturierte Daten“, Ziele und Adressaten, Zusammenfassung des Workshops und Einführung TEF
- Hauptkapitel untergliedert nach den Fehlerdimensionen des TEF: Datengenerierung, Datenaufbereitung, Datenanalyse
  - Am Ende der Kapitel Empfehlungen
- Abschluss mit Ausblick zu offenen Fragen und Herausforderungen bei der Forschung mit unstrukturierten Daten

Inhaltsverzeichnis

Herausforderungen bei unstrukturierten Daten

- Einleitung
- Prof. Dr. Theresa Gessler, Europa-Universität Viadrina Frankfurt (Oder)
  - Herausforderungen bei der Nutzung unstrukturierter Daten: Eine politikwissenschaftliche Perspektive
- Prof. Dr. Florian Keusch, Universität Mannheim
  - Sammlung digitaler Verhaltensdaten mittels Smartphone: Anwendungen, Chancen und Herausforderungen
- Dr. Katrin Weller, GESIS – Leibniz-Institut für Sozialwissenschaften
  - „Vergänglichkeit" als Herausforderung im Umgang mit Daten aus Online-Plattformen
- Abschlussdiskussion

Herausforderungen bei unstrukturierten Daten

- **Forschung**
  - Repräsentativität, Coverage (Smartphone), (Nicht-)Teilnahme
  - Vergänglichkeit von Online-Plattform-Daten und mangelnde Dokumentation
    - Veränderungen der Plattform Affordances
    - Veränderungen des Nutzungsverhaltens
    - Veränderungen der Inhalte
  - Unklarer Prozess der Datengenerierung
  - Probleme mit Messung und Konzeptualisierung, Messqualität
- **Datenzugang, -archivierung und -nachnutzung**
- **Methodenausbildung**