gesis

Leibniz-Institut
für Sozialwissenschaften

More data, more problems?

Chancen und Herausforderungen der datafizierten Gesellschaft für die sozialwissenschaftliche Forschung

Johannes Breuer, Sebastian Stier, Pascal Siegers

Mitglied der Leibniz-Gemeinschaft

Hintergrund

- Nutzer_innen digitaler Medien produzieren immer mehr Daten
- solche digitalen (Verhaltens-) Spurdaten sind auch für die sozialwissenschaftliche Forschung äußerst interessant
  - Webtracking-Daten
  - Social-Media-Daten
  - …

Selbstauskünfte & digitale Spurdaten

- Selbstauskünfte zu Mediennutzung oft nicht verlässlich (siehe z.B. Scharkow, 2016; Araujo et al., 2017)
  - soziale Erwünschtheit
  - Schwierigkeiten bei Erinnerung oder mentaler Berechnung (z.B. Fragen zu spezifischen/seltenen Nutzungsweisen oder längeren Zeiträumen)
- digitale Spurdaten alleine oft problematisch für sozialwissenschaftliche Fragestellungen
  - fehlende Informationen zu Personen
  - fehlende relevante abhängige Variable (z.B. Einstellungen oder Verhalten)
  - u.U. forschungsethisch schwierig: z.B. kein Informed Consent

Data Linking

- Kombination von Befragungs- und digitalen Spurdaten als Antwort auf die Limitationen der beiden Datentypen (Stier et al., 2019)
- verschiedene Wege der Verknüpfung
- spezifische methodische und ethische Herausforderungen für solche verbundenen Daten

Linking-Typologie

Ex Ante	Ex Post

Aggregatebene
• Analyse von Audience Overlap (z.B. Mukerjee et al., 2018; Nelson & Webster, 2017)
• Analyse aggregierter Publikums-/Nutzer_innenstatistiken (z.B. Nelson & Webster, 2017)
	Aggregatebene
	• Verknüpfung auf von Ebene von…
	  • Zeit (z.B. Mellon, 2014; O'Connor et al., 2010; Stier et al., 2018)
	  • Themen (z.B. Pasek et al., 2019)
	  • Geographie (z.B. Beauchamp, 2017)

	Öffentliche Akteure • z.B. Politiker_innen oder Organisationen (siehe z.B. Karlsen & Enjolras, 2016; Quinlan et al., 2017)

Individualebene
• Informed Consent in Befragung für Erhebung der digitalen Spurdaten:
  • Webseitenbesuche (z.B. Guess, 2015; Jürgens et al., 2019; Möller et al., 2019; Vraga & Tully, 2018)
  • Smartphone-Daten (z.B. Boase & Ling, 2013; Jürgens et al., 2019; Kreuter et al., 2019)
  • Sensordaten (z.B. Génois, Zens, Lechner, Rammstedt, & Strohmaier, 2019)
	Individualebene
	• Informed Consent in Befragung für retrospektive Erhebung:
	  • über Social-Media-APIs (z.B. Al Baghal et al., 2019; Haenschen, 2019; Hofstra, Corten, van Tubergen, & Ellison, 2017; Hopp, Vargo, Dixon, & Thain, 2018; Vaccari et al., 2015; Wells & Thorson, 2017)
	  • über Datenspende; z.B. persönliches Twitter-Archiv oder Export eigener Facebook-Daten (siehe z.B. Thorson et al., 2019)

Quelle: Stier et al., 2019

Beispiele spezieller Herausforderungen

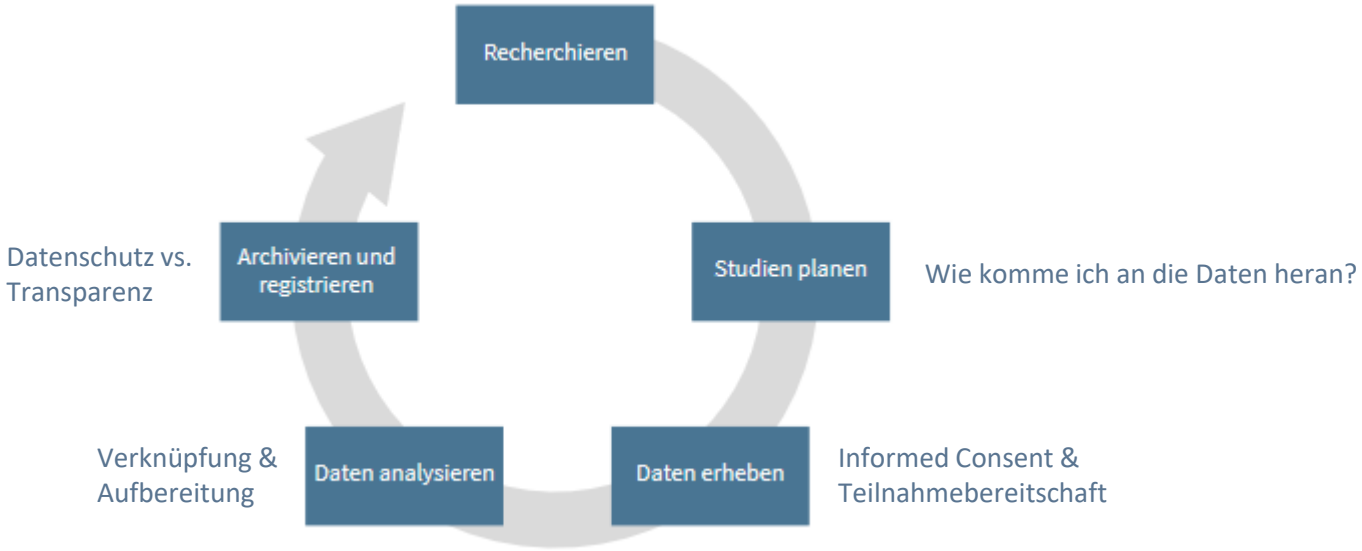
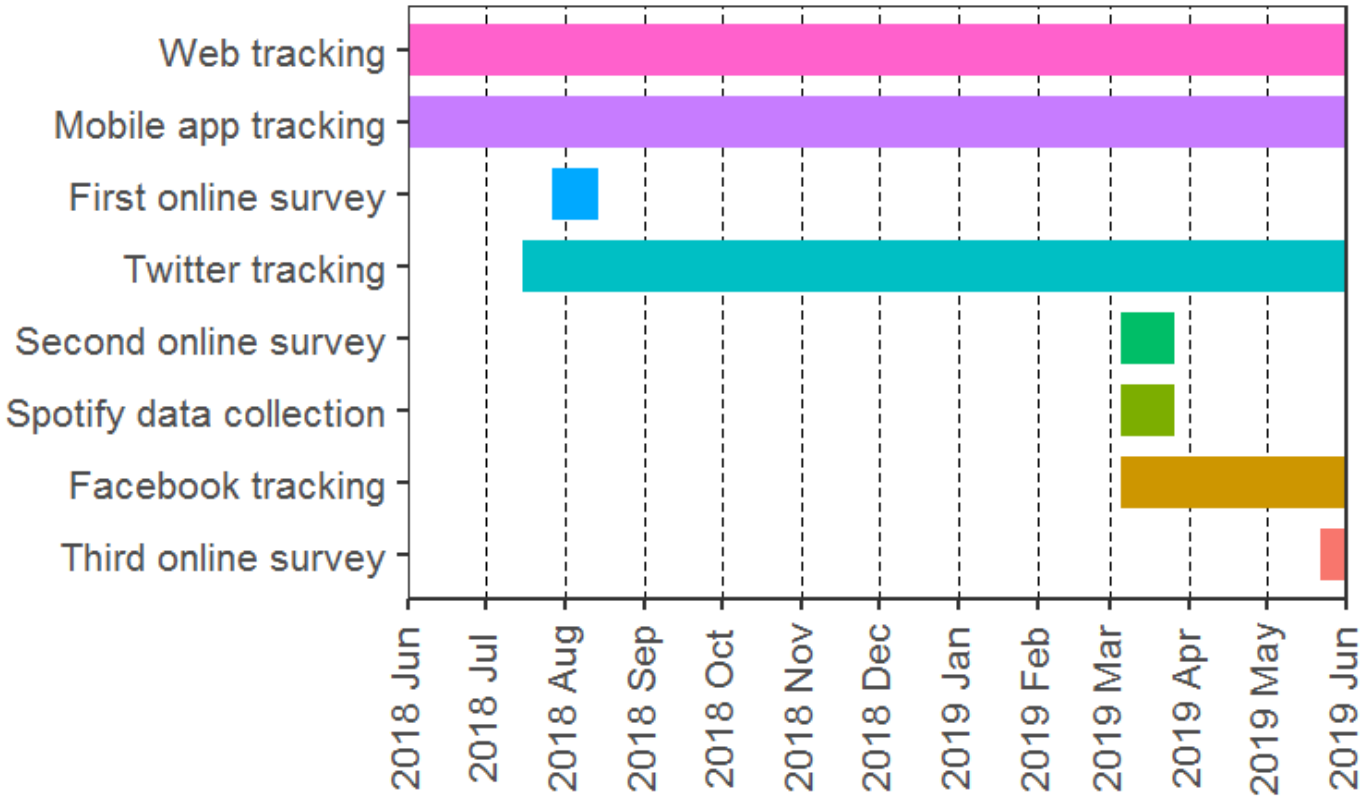


Abbildung: https://www.gesis.org/angebot

Unser Projekt

- *respondi*-Web-Tracking-Panel
  - Daten für 1 Jahr (Juni 2018 – Mai 2019)
  - ~ N = 2000 Teilnehmer_innen pro Monat
  - insg. ca. 94 Mio. Datenpunkte
  - für Teil der Teilnehmer_innen: zusätzlich Tracking der mobilen App-Nutzung
- 3 Online-Befragungen der Panelist_innen
  - thematischer Fokus: Mediennutzung & Politik
- Social-Media-Daten
  - Twitter
  - Facebook
  - Spotify

Im Projekt erhobene Datentypen

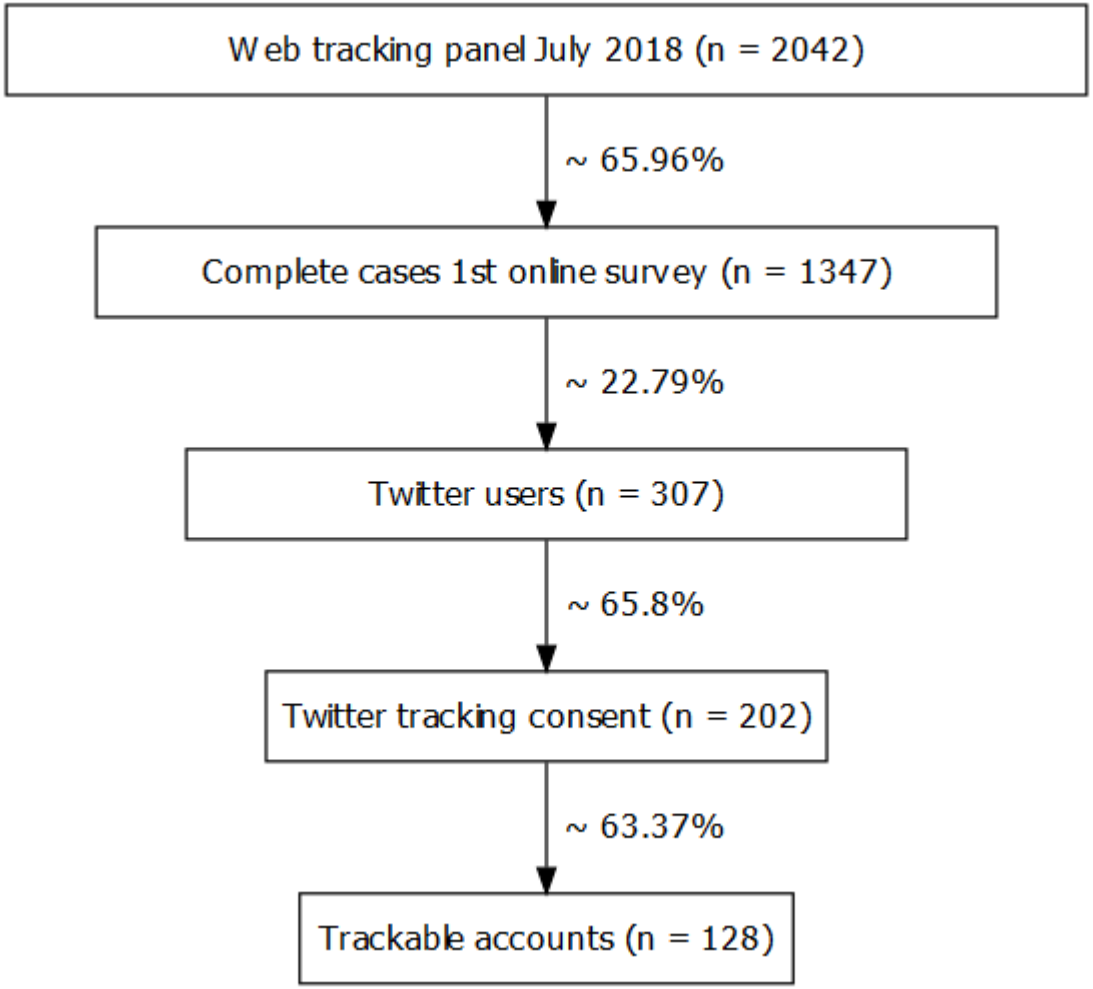


Social-Media-Daten

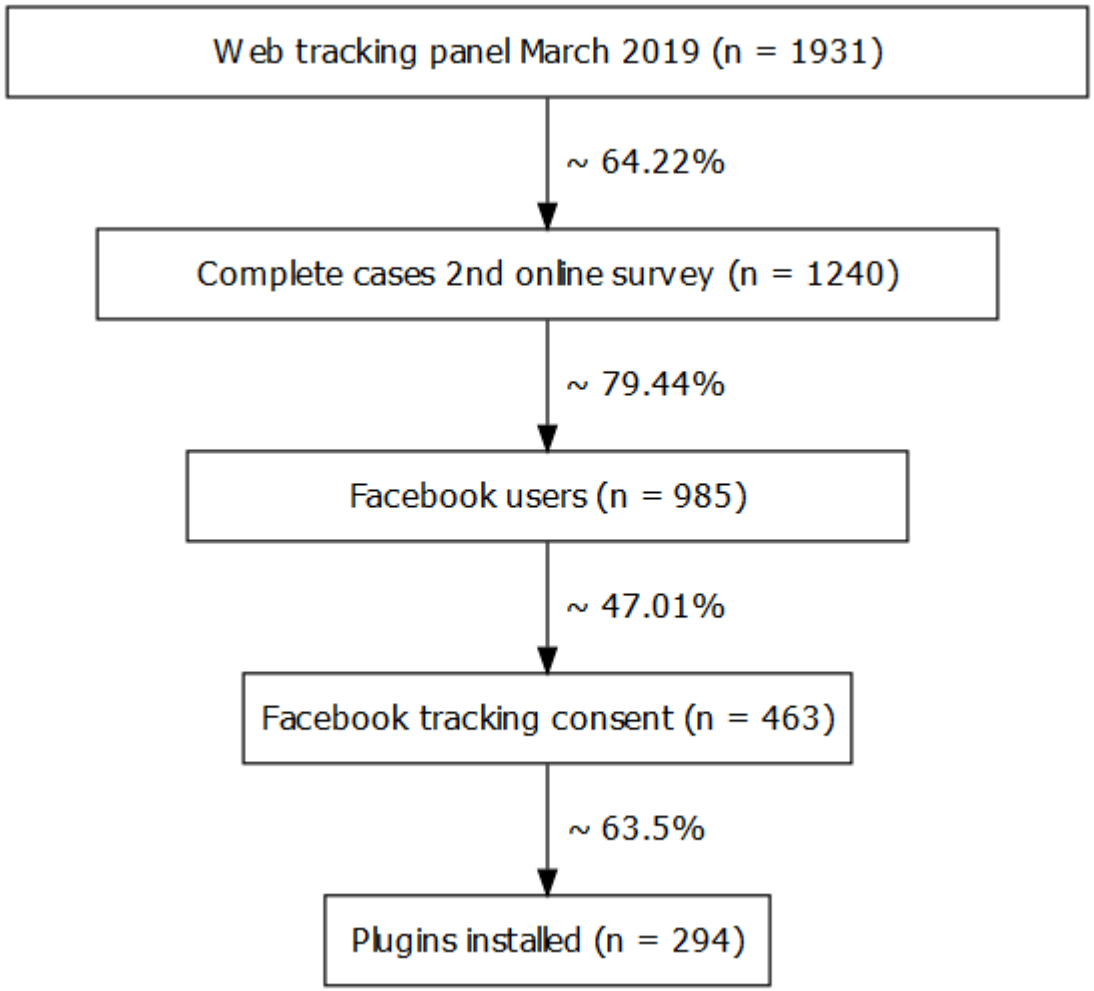
- Twitter
  - kontinuierliches Tracking via Streaming API
- Facebook
  - Browser-Plugin (Haim & Nienierza, 2019)
    - für Firefox & Chrome
    - sammelt öffentliche Posts (+ einige Metadaten zu diesen) aus dem Feed der Nutzer_innen
- Spotify
  - Web-App von der KU Leuven
    - sammelt die 50 zuletzt abgespielten Songs, Playlists & Präferenzen

- Für alle Daten:
  - kurzer Informed Consent in Online-Befragung
  - ausführlichere Informationen (zu Datenschutz etc.) auf Webseite (im Fragebogen verlinkt)
  - Informed Consent & Privacy Information von Al Bahgal et al. (2019) übernommen, übersetzt und angepasst

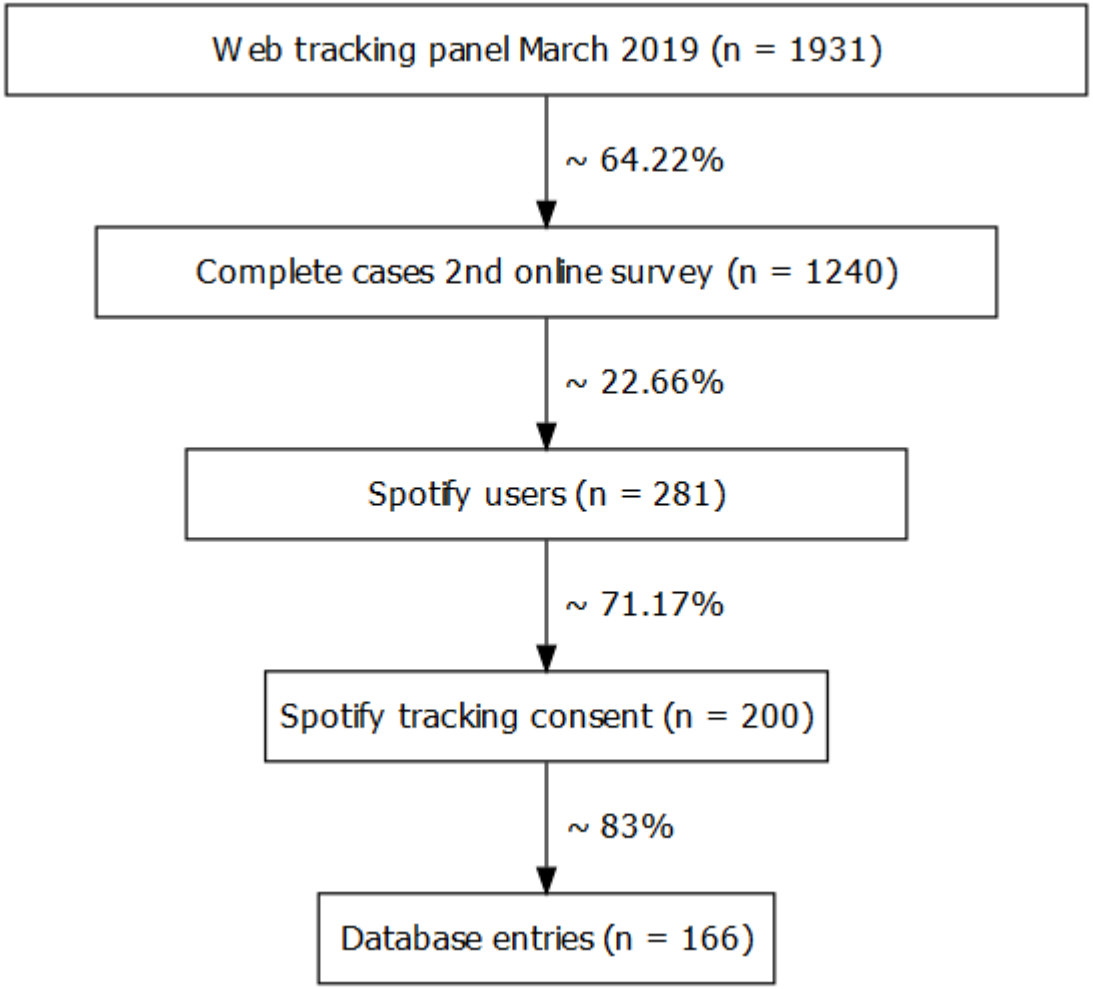
Twitter-Daten



Facebook-Daten



Spotify-Daten



Nächste Schritte & offene Fragen

- systematischer Bias in den Dropout-Stufen?
- Qualität/Nutzbarkeit der Social-Media-Daten?
  - Nutzungsaktivität quantitativ -> Datenpunkte pro Person
  - Nutzungsaktivität qualitativ -> z.B. passive vs. aktive Nutzung bei Twitter; Facebook als Nachrichtenquelle?
- geplante Lösung für Data Sharing
  - frei zugängliche aggregierte Daten über GESIS datorium
  - Individualdaten über GESIS Secure Data Center

Vor- und Nachteile unseres Ansatzes

Vorteile	Nachteile
• Daten auf individueller Ebene • große & heterogene Stichprobe • große Bandbreite an Daten • Informierte Einwilligung der Teilnehmer_innen • unaufwändiger Zugang zu Webtracking-Daten • Facebook-Daten nicht über API	• Bias(es) in der Stichprobe • Kosten für Webtracking-Daten • Twitter & Spotify über API • Risiko der Veränderung oder Schließung von APIs (Bruns, 2019; Freelon, 2018; Puschmann, 2019) • Veränderungen in der Struktur des Facebook-Feeds u.U. problematisch für Browser-Plugin

Alternative Datenzugänge

Datenart	Methode	Vorteile	Nachteile
Web- und mobile App-Nutzung	Eigene Software (+ eigenes Panel)	• Mehr Kontrolle über Rekrutierung und Erhebung	• Hoher Aufwand (Erstellung der Software/App + Support; Rekrutierung der Teilnehmer_innen)
Social-Media-Daten	„Datenspende“ (Nutzer_innen exportieren eigene Daten und teilen diese mit den Forscher_innen; siehe Halavais, 2019; Thorson et al., 2019)	• Direkte Einbindung der Teilnehmer_innen • Transparenz	• Aufwand für Teilnehmer_innen (Export der Daten nicht selbsterklärend und i.d.R. nicht direkt möglich) • Lösung zum Teilen der Daten (inkl. Anonymisierung) nötig

Zusätzlich Variation der Rekrutierungswege/-reihenfolge möglich:

- z.B. Befragung (-> Webtracking) -> Social Media
- Reihenfolge hat Einfluss auf Zusammensetzung der Stichprobe & Biases

Zusammenfassung

- Befragungs- und digitale Spurdaten haben jeweils eigene Limitationen
- Verknüpfung dieser beiden Datentypen ist vielversprechend für die sozialwissenschaftliche Forschung
- Verknüpfung bringt spezifische Herausforderungen für alle Phasen des Forschungsdatenzyklus mit sich
- es gibt verschiedene Möglichkeiten, die Daten zu erheben und sie zu verknüpfen
- richtige Auswahl der Methoden hängt von der Fragestellung sowie den verfügbaren Ressourcen eines Projekts ab