gesis Leibniz-Institut für Sozialwissenschaften

Verknüpfung von digitalen Spurdaten & Umfragen

Wieso, wie und was ist möglich?

Johannes Breuer

25.01.2023

Leibniz Gemeinschaft

Vorstellung

- Senior Researcher im Team Data Augmentation, Abteilung Survey Data Curation, GESIS – Leibniz-Institut für Sozialwissenschaften (Köln) & Co-Leiter des Teams Research Data & Methods am Center for Advanced Internet Studies (CAIS, Bochum)
  - Arbeitsschwerpunkte: Nutzung digitaler Spurdaten in der sozialwissenschaftlichen Forschung & Verknüpfung mit Befragungsdaten
- Weitere Forschungsschwerpunkte:
  - Nutzung und Wirkung digitaler Medien
  - Computational Methods
  - Open Science
- Disziplinärer Hintergrund: (Medien-)Psychologie & Kommunikationswissenschaft
- Mehr Informationen: https://www.johannesbreuer.com/

Was ist/macht GESIS?

gesis Leibniz-Institut für Sozialwissenschaften

- Institut der Leibniz-Gemeinschaft
  - Leibniz-Gemeinschaft (97 Institute) = eine der vier großen außeruniversitären Forschungsorganisationen in Deutschland neben Fraunhofer-Gesellschaft, Helmholtz-Gemeinschaft, Max-Planck-Gesellschaft
- GESIS-Standorte: Köln & Mannheim
- größte Infrastruktureinrichtung für Sozialwissenschaften in Europa (Forschung & Service)
- großes Portfolio an Angeboten und Services für Forschende (richtet sich auch an Studierende, z.B. die Trainings- und Weiterbildungsangebote im Bereich Methoden)
- Neuer Sondertatbestand Digitale Verhaltensdaten
- GESIS @ Social Media
  - Twitter
  - Mastodon
  - Facebook
  - Instagram
  - YouTube
- GESIS betreibt auch einen eigenen Podcast sowie einen Blog

gesis Leibniz-Institut für Sozialwissenschaften

Was ist/macht das CAIS?

CAIS RESEARCH FOR THE DIGITAL AGE

- Standort: Bochum
- Gegründet 2017 als Wissenschaftskolleg
- Inzwischen fortlaufender Ausbau zu einem Institut für Digitalisierungsforschung mit eigenen Forschungsprogrammen aus verschiedenen Disziplinen und zu unterschiedlichen Digitalisierungsthemen (z.B. Partizipation, Lernen, KI)
- Verschiedene Unis & Forschungsinstitute aus dem NRW sind Gesellschafter
- CAIS @ Social Media
  - Twitter
  - Facebook
  - Instagram
  - YouTube
- CAIS-Podcast

Inhalt

1. ***Warum*** sollte man Umfragen & digitale Spurdaten verknüpfen?
2. ***Wie*** kann man Befragungs- und digitale Spurdaten verknüpfen?
3. ***Was*** kann man mit solchen verknüpften Daten machen?
4. Welche ***Herausforderungen*** gibt es bei der Verknüpfung?
5. Welche ***Lösungsansätze*** gibt es dafür?

Terminologie

- Wenn es um die Verknüpfung von (Forschungs-)Daten geht, findet man in der Literatur verschiedene Begriffe:
  - Data Linking/Linkage
  - Record Linkage
  - Linkage Analysis

- Manchmal unterscheiden diese Begriffe sich in ihrer Bedeutung, aber bei allen geht es darum, unterschiedliche Daten bzw. Datentypen und -quellen miteinander zu verbinden
- Ich selbst präferiere und verwende i.d.R. den Begriff Data Linking (dieser ist auch bei GESIS etabliert)

Terminologie

- Auch für den Datentyp, der in dieser Veranstaltung näher in den Blick genommen wird, gibt es verschiedene Bezeichnungen:
  - Digital Behavioral Data/digitale Verhaltensdaten (DVD)
  - Digital Trace Data/digitale Spurdaten
  - Social Web Data
  - Big Data

- Bei GESIS ist der Begriff digitale Verhaltensdaten (DVD) etabliert
- Ich selbst nutze meistens den Begriff digitale Spurdaten und werde diesen und DVD heute synonym verwenden

Warum Umfragen & DVD verknüpfen?

TL;DR-Antwort

- Digitale Spurdaten & Befragungsdaten haben jeweils spezifische Stärken und Limitationen
- Durch die Verknüpfung dieser beiden Datentypen kann man die jeweiligen Stärken kombinieren (und den Limitationen ein Stück weit begegnen)
- Mit den verknüpften Daten lassen sich neue Forschungsfragen beantworten bzw. neue Antworten auf bestehende Fragen finden

- siehe Stier et al., 2020

Befragungsdaten

Stärken	Limitationen
• Erlauben die Erhebung detaillierter Informationen über Eigenschaften von Personen (z.B. Soziodemographie)	• Mögliche Validitäts- und Reliabilitätsprobleme bei Selbstauskünften (z.B. zur Mediennutzung)
• Können eine Vielzahl von Dimensionen erfassen: Meinungen, Einstellungen, Verhalten	• Mögliche Verzerrungen (Biases) durch soziale Erwünschtheit
• Können Offline- und Online-Verhalten erfassen	• Verhalten wird i.d.R. retrospektiv abgefragt (→ Erinnerungsschwierigkeiten)
• Wahrscheinlichkeitsstichproben (Probability Samples) können gezogen werden	• Sinkende Antwortraten (insb. für Telefonbefragungen)

Digitale Spurdaten

Stärken	Limitationen
• Direkte Erfassung von Verhalten	• Nur begrenzte Informationen über die datengenerierenden Individuen (z.B. im Hinblick auf ihre Soziodemographie)
• Hohe zeitliche Auflösung	• Keine direkte Messung von Einstellungen
• Daten werden in großer Menge (Volume) und Geschwindigkeit (Velocity) generiert	• Keine bzw. kaum Informationen über Offline-Aktivitäten
• i.d.R. geringere Kosten als für Surveys	• Unsicherheiten/Unwägbarkeiten im Datenzugang (z.B. im Falle der Nutzung von APIs)

Wie kann man Umfragen & DVD verknüpfen?

- Drei Dimensionen, auf denen sich Ansätze zur Verknüpfung von Umfragedaten und DVD im Wesentlichen unterscheiden können:
  - Zeitliche Abfolge der Datensammlung und -verknüpfung
  - Sicherheit/Gewissheit der Verknüpfung
  - Ebene der Verknüpfung

Verknüpfungsarten

- Zeitliche Abfolge der Datensammlung und -verknüpfung
  - **Ex-ante Linking**: Daten werden spezifisch für die Verknüpfung in der jeweiligen Studie bzw. im jeweiligen Projekt gesammelt
  - Ex-Post Linking: Mindestens einer der zu verknüpfenden Datensätze existierte bereits vorher (z.B. Befragungsdaten aus großen Umfrageprogrammen oder abgeschlossene und geteilte Sammlungen von Social-Media-Daten) → (partielle) Sekundäranalyse

Verknüpfungsarten

- Sicherheit/Gewissheit der Verknüpfung
  - Probabilistic Linking/Linkage: Keine eindeutige Zuordnung zwischen Beobachtungseinheiten[1] in den unterschiedlichen Datenquellen möglich, daher Anwendung von Schätzverfahren zur Identifikation von Übereinstimmungen (Matches)
  - **Deterministic Linking**: Eindeutige Zuordnung zwischen Beobachtungseinheiten in den Datenquellen möglich (z.B. durch eine geteilte ID-Variable)

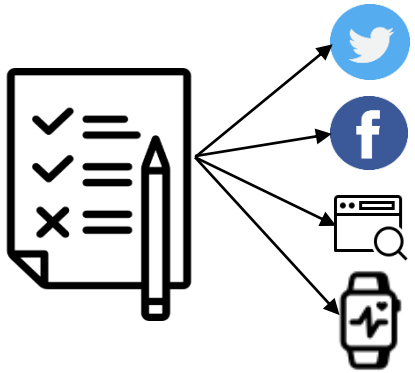
[1] In den folgenden Beispielen sind die Beobachtungseinheiten Individuen, aber dies könnten z.B. auch Organisationen sein (Parteien, Unternehmen, NGOs, etc.)

Verknüpfungsarten

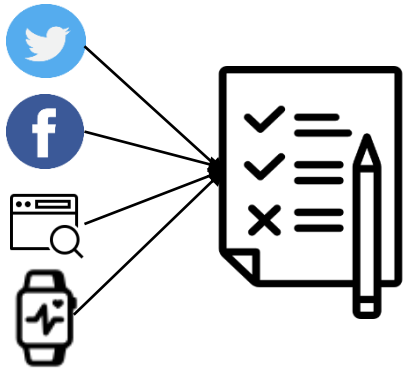
- Ebene der Verknüpfung
  - Aggregatebene: Z.B. für geographische Einheiten (z.B. Länder/Regionen) und/oder bestimmte Zeiträume (oder auch spezifische Themen)
  - **Individualebene**: Daten für/über einzelne Nutzer:innen bzw. Teilnehmer:innen

Reihenfolge der Datensammlung

- NB: Hier und im Folgenden fokussieren wir uns auf den Fall des **deterministischen ex-ante Linking auf Individualebene**



Survey first: Anfrage zur Verknüpfung in Befragung



DVD first: Einladung zum Survey über Plattform/Website/App

- Beide Varianten bergen das Risiko bestimmter systematischer Verzerrungen (Biases): z.B. Opt-In Bias(es), Unterschiede zwischen Nutzer:innen bestimmter Plattformen (und der Allgemeinbevölkerung)…
  - Zum Thema DVD & Bias, siehe Sen et al., 2021

gesis Leibniz-Institut für Sozialwissenschaften

Der ideale verknüpfte Datensatz?

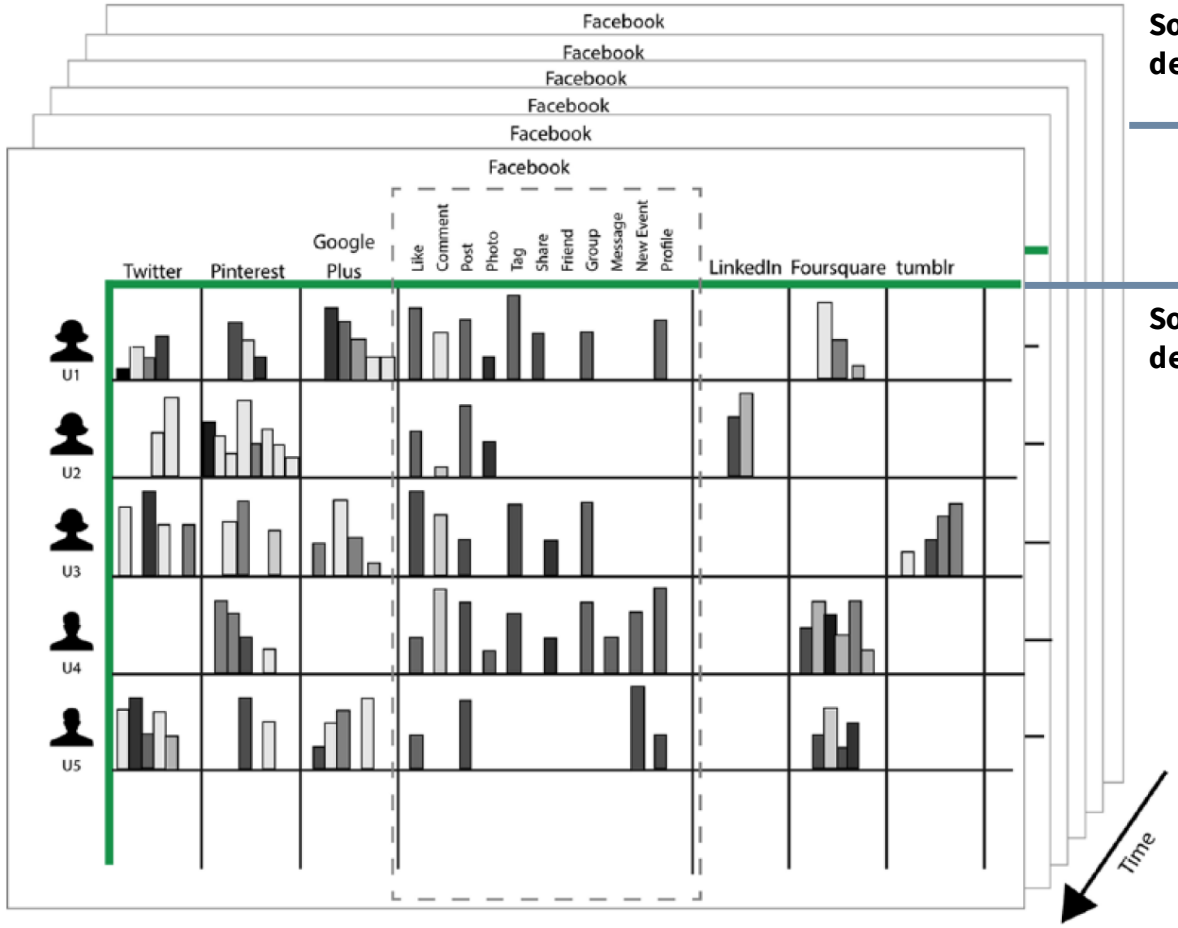


Abbildung basierend auf Resnick et al. (2015)

Sozio-demographie	Offline Netzwerke	(Politische) Einstellungen	Psychologische Variablen

Sozio-demographie	Offline Netzwerke	(Politische) Einstellungen	Psychologische Variablen

Umfragen

gesis Leibniz-Institut für Sozialwissenschaften

Was ist mit verknüpften Umfragen & DVD möglich?

Article

Self-Reported Versus Digitally Recorded: Measuring Political Activity on Facebook

Social Science Computer Review
1-17

DOI: 10.1177/0894439318813586
journals.sagepub.com/home/ssc
SAGE

Katherine Haenschen[1]

Abstract
Facebook has been credited with expanding political activity by simultaneously lowering barriers to participation and creating new ways to engage. However, many of these findings rely on subjects' abilities to accurately report their Facebook use and political activity on the platform. This study combines survey responses and digital trace data from 828 American adults to determine whether subjects over- or underreport a range of political activities on Facebook, including whether they like political pages or share news links. The results show that individuals underestimate their frequency of status posting and overestimate their frequency of sharing news links on Facebook. Political interest is associated with a decrease in underreporting several political activities, while increasing the likelihood of overreporting the frequency of sharing news links. Furthermore, political interest serves a moderating effect, improving self-reports for high-volume users. The findings suggest that political interest not only predicts political activity but also shapes awareness of that activity and improves self-reports among heavy users.

Keywords
political participation, political interest, social media, Facebook, digital trace data

Article

Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept

Social Science Computer Review
1-16

DOI: 10.1177/0894439318814241
journals.sagepub.com/home/ssc
SAGE

Toby Hopp[1], **Chris J. Vargo**[1], **Lucas Dixon**[2], **and Nithum Thain**[2]

Abstract
This study correlated self-report and trace data measures of political incivility. Specifically, we asked respondents to provide estimates of the degree to which they engage in uncivil political communication online. These estimates were then compared to computational measures of uncivil social media discussion behavior. The results indicated that those who self-disclose uncivil online behavior also tend to generate content on social media that is uncivil as identified by Google's Perspective application programming interface. Taken as a whole, this work suggests that combining self-report and behavioral trace data may be a fruitful means of developing multimethod measures of complex communication behaviors.

Keywords
incivility, political discussion, toxicity, survey, computational social sciences

Article

Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces?

Social Science Computer Review
1-18

DOI: 10.1177/0894439318822007
journals.sagepub.com/home/ssc
SAGE

Josh Pasek[1], **Colleen A. McClain**[1], **Frank Newport**[2], **and Stephanie Marken**[2]

Abstract
Researchers hoping to make inferences about social phenomena using social media data need to answer two critical questions: What is it that a given social media metric tells us? And who does it tell us about? Drawing from prior work on these questions, we examine whether Twitter sentiment about Barack Obama tells us about Americans' attitudes toward the president, the attitudes of particular subsets of individuals, or something else entirely. Specifically, using large-scale survey data, this study assesses how patterns of approval among population subgroups compare to tweets about the president. The findings paint a complex picture of the utility of digital traces. Although attention to subgroups improves the extent to which survey and Twitter data can yield similar conclusions, the results also indicate that sentiment surrounding tweets about the president is no proxy for presidential approval. Instead, after adjusting for demographics, these two metrics tell similar macroscale, long-term stories about presidential approval but very different stories at a more granular level and over shorter time periods.

Keywords
Twitter sentiment, presidential approval, demographics, trends over time

Artikel aus dem Special Issue im Journal *Social Science Computer Review zum Thema* „Integrating Survey Data and Digital Trace Data",
(Guest Editors: Sebastian Stier, Johannes Breuer, Pascal Siegers, & Kjerstin Thorson)

Was ist mit verknüpften Umfragen & DVD möglich?

Article

Explaining Online News Engagement Based on Browsing Behavior: Creatures of Habit?

Social Science Computer Review
1-17

DOI: 10.1177/0894439319828012
journals.sagepub.com/home/ssc
SAGE

Judith Möller[1], Robbert Nicolai van de Velde[1], Lisa Merten[2], and Cornelius Puschmann[2]

Abstract
Understanding how citizens keep themselves informed about current affairs is crucial for a functioning democracy. Extant research suggests that in an increasingly fragmented digital news environment, search engines and social media platforms promote more incidental, but potentially more shallow modes of engagement with news compared to the act of routinely accessing a news organization's website. In this study, we examine classic predictors of news consumption to explain the preference for three modes of news engagement in online tracking data: routine news use, news use triggered by social media, and news use as part of a general search for information. In pursuit of this aim, we make use of a unique data set that combines tracking data with survey data. Our findings show differences in predictors between preference for regular (direct) engagement, general search-driven, and social media–driven modes of news engagement. In describing behavioral differences in news consumption patterns, we demonstrate a clear need for further analysis of behavioral tracking data in relation to self-reported measures in order to further qualify differences in modes of news engagement.

Keywords
news use, tracking data, survey data, social media, information search

Article

Who Is Exposed to News? It Depends on How You Measure: Examining Self-Reported Versus Behavioral News Exposure Measures

Social Science Computer Review
1-17

DOI: 10.1177/0894439318812050
journals.sagepub.com/home/ssc
SAGE

Emily K. Vraga[1] and Melissa Tully[2]

Abstract
Despite the importance of news exposure to political outcomes, news consumption is notoriously difficult to measure, and misreporting news exposure is common. In this study, we compare participants' news behaviors measured on a news aggregator website with their self-reported story selection immediately after exposure. We find that both individual and contextual characteristics—especially the presence of political cues in news headlines—influence reporting of news story selection. As a result, the news audience profiles differ using self-reported versus behavioral measures, creating two different pictures of news exposure. More attention is needed to improve news measurement strategies to address misreporting and to improve the accuracy of news audience profiles.

Keywords
news exposure, self-report measures, website analytics, news audiences, misreporting

Artikel aus dem Special Issue im Journal *Social Science Computer Review zum Thema* „Integrating Survey Data and Digital Trace Data", (Guest Editors: Sebastian Stier, Johannes Breuer, Pascal Siegers, & Kjerstin Thorson)

Beispiele aus unserer eigenen Forschung

SM+S social media + society

Article

Social Media + Society
July-September 2021: 1–11

DOI: 10.1177/20563051211033820
journals.sagepub.com/home/sms
SAGE

Do News Actually "Find Me"? Using Digital Behavioral Data to Study the News-Finds-Me Phenomenon

Mario Haim[1], Johannes Breuer[2], and Sebastian Stier[2]

Abstract

Research on news exposure has shown that while political knowledge and interest largely determine the degree of active engagement with online news, some people are generally less willing to invest into actively staying informed. Instead, these people report to pursue a passive mode of relying on specific sources, such as social media, based on the belief that "news finds me" (NFM). Notably, the three dimensions of NFM—feeling informed, relying on peers, and not actively seeking news—combine intentions and perceptions related to news use. Understanding NFM perceptions, hence, requires an analytical distinction between active and passive modes of news use as well as reliable measures of (different types of) news exposure. We contribute to this field by combining a survey, tracked web-browsing data, and tracked Facebook data to investigate the relationship between NFM perceptions and exposure to online news, also taking into account political knowledge and interest as traditional predictors of active news use. Our results show that both political knowledge and interest are associated with more news exposure via web browsers and that political knowledge—but not political interest—is also associated with more news in people's Facebook feeds. Compared with the NFM dimensions, political knowledge and interest are stronger predictors of online news exposure in our study. Taken together, the novel combination of Facebook and web tracking data provides evidence that online news exposure is shaped by a confluence of traditional factors and more diffuse interpersonal processes.

Keywords

news finds me, social media, web tracking, diffusion, media use, news, Facebook

American Political Science Review (2022) 116, 2, 768–774
doi:10.1017/S0003055421001222

Letter

Post Post-Broadcast Democracy? News Exposure in the Age of Online Intermediaries

SEBASTIAN STIER *GESIS – Leibniz Institute for the Social Sciences, Germany*
FRANK MANGOLD *University of Hohenheim, Germany*
MICHAEL SCHARKOW *Johannes Gutenberg University Mainz, Germany*
JOHANNES BREUER *GESIS – Leibniz Institute for the Social Sciences, Germany, and Center for Advanced Internet Studies, Germany*

Online intermediaries such as social network sites or search engines are playing an increasingly central role in democracy by acting as mediators between information producers and citizens. Academic and public commentators have raised persistent concerns that algorithmic recommender systems would negatively affect the provision of political information by tailoring content to the predispositions and entertainment preferences of users. At the same time, recent research indicates that intermediaries foster exposure to news that people would not use as part of their regular media diets. This study investigates these unresolved questions by combining the web browsing histories and survey responses of more than 7,000 participants from six major democracies. The analysis shows that despite generally low levels of news use, using online intermediaries fosters exposure to nonpolitical and political news across countries and personal characteristics. The findings have implications for scholarly and public debates on the challenges that high-choice digital media environments pose to democracy

gesis Leibniz-Institut für Sozialwissenschaften

Beispiele aus unserer eigenen Forschung

POLITICAL RESEARCH EXCHANGE
2022, VOL. 4,
https://doi.org/10.1080/2474736X.2022.2135451

ecpr Routledge Taylor & Francis Group

Research Article

OPEN ACCESS Check for updates

The role of the information environment during the first COVID-19 wave in Germany

Sebastian Stier [a], Bernd Weiß [a], Timo Hartmann[a], Fabian Flöck[a], Johannes Breuer [a,b], Ines Schaurer[a] and Mirjan Kummerow[a]

[a]GESIS – Leibniz Institute for the Social Sciences, Mannheim/Cologne, Germany; [b]Center for Advanced Internet Studies (CAIS), Bochum, Germany

ABSTRACT

The COVID-19 pandemic has been accompanied by intense debates about the role of the information environment. On the one hand, citizens learn from public information campaigns and news coverage and supposedly adjust their behaviours accordingly; on the other, there are fears of widespread misinformation and its detrimental effects. Analyzing the posts of the most important German information providers published via Facebook, this paper first identifies a uniform salience of subtopics related to COVID-19 across different types of information sources that generally emphasized the threats to public health. Next, using a large survey conducted with German residents during the first COVID-19 wave in March 2020 we investigate how information exposure relates to perceptions, attitudes and behaviours concerning the pandemic. Regression analyses show that getting COVID-19-related information from a multitude of sources has a statistically significant and positive relationship with public health outcomes. These findings are consistent even across the ideological left/right spectrum and party preferences. These consistent correlational results demonstrate that during the first wave of COVID-19, a uniform information environment went hand in hand with a cautious public and widely accepted mitigation measures. Nonetheless, we discuss these findings against the backdrop of an increased politicization of public-health measures during later COVID-19 waves.

ARTICLE HISTORY
Received 9 November 2021
Accepted 6 October 2022

KEYWORDS
COVID-19; information exposure; infodemic; Facebook; public broadcasting

A reevaluation of online pornography use in Germany using a combination of web tracking and survey data

Maximilian T. P. von Andrian-Werburg[1], Pascal Siegers[2], Johannes Breuer[2,3]

Institute Human-Computer-Media, University of Würzburg[1]

GESIS – Leibniz Institute for the Social Sciences[2]

Center for Advanced Internet Studies (CAIS)[3]

Several researchers have questioned the reliability of pornography research's findings as they may, e.g., be biased by social desirability. Following a recent call to make use of more reliable data sources, we conducted two studies to investigate patterns and differences as well as predictors of online pornography use (OPU). In our first study, we used data from a large-scale German online web tracking panel (N = 3018, website visits on domain level) gathered from June 2018 to June 2019. We looked at group differences as well as temporal trends. Overall, our results confirm existing findings from questionnaire-based research related to sex and age differences. Our data also shows temporal patterns, which indicate that – for the majority of users – OPU constitutes a form of leisure time activity that competes with other spare time activities but also family and social obligations (we found the lowest amount of pornography use on Christmas). In our second study, we linked the web tracking to data from an online survey (N = 1315) to reassess the relevance of various predictors of OPU that have been identified in previous research. As in Study 1, our results mostly echo previous findings. Online pornography is used more by males and younger individuals, while relationship, sexism, and social dominance orientation are not associated with OPU. However, we do find differences in OPU between members of different religious communities. With our two studies, we were able to confirm some key findings on OPU from previous questionnaire-based research using web tracking data, while also being able to engage in more finely-grained analyses of usage patterns.

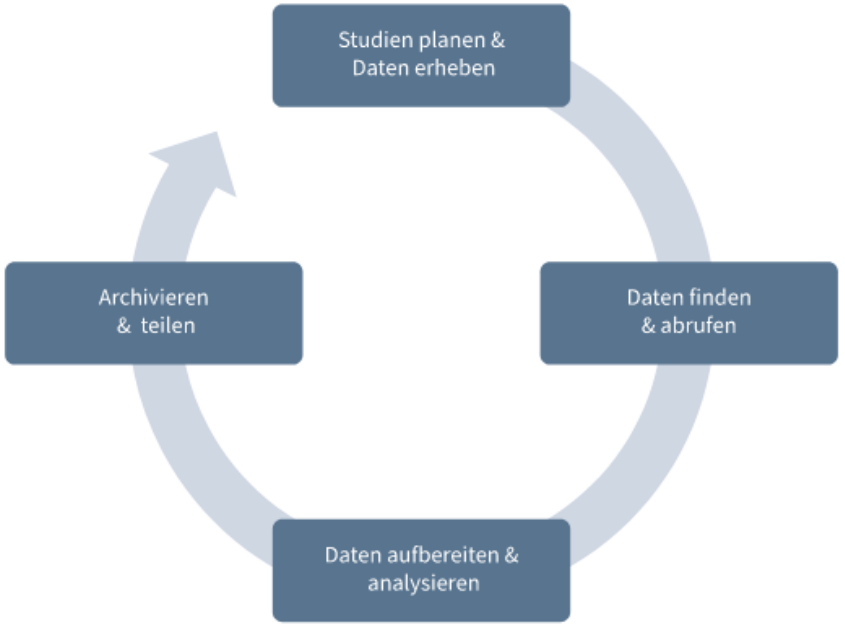
Key words – online pornography use, web tracking data, religiosity, sexism, social dominance orientation

https://psyarxiv.com/ehqgv/

Herausforderungen bei der Verknüpfung

- Verschiedene Arten von Herausforderungen bei der Verknüpfung von Umfragen und DVD:
  - Forschungspraktische
  - Rechtliche & ethische

- Herausforderungen in allen Phasen des Forschungsdatenzyklus bzw. in allen Forschungsphasen



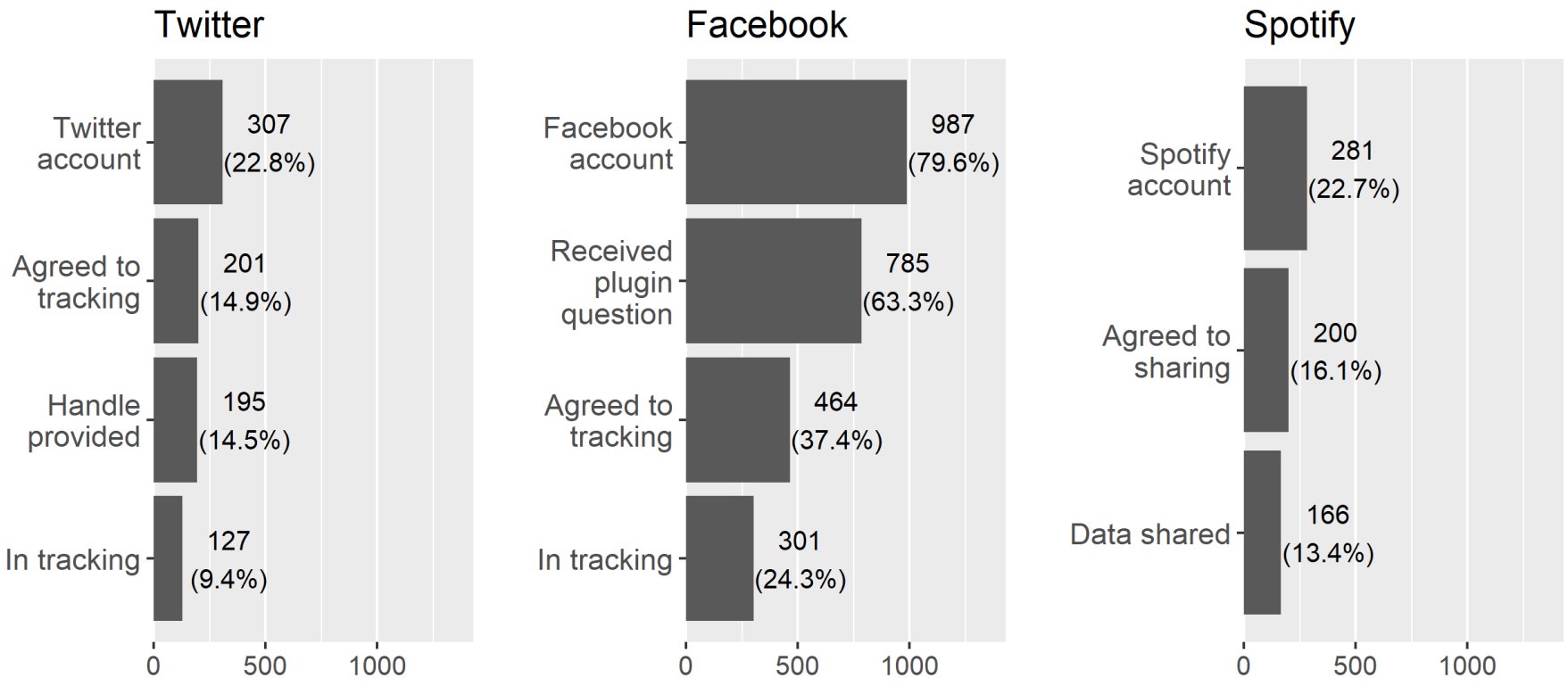
Quelle der Abbildung: https://www.gesis.org/angebot

Forschungspraktische Herausforderungen

- **Datenzugang** für DVD
  - API (siehe z.B. Bruns, 2019; Freelon, 2018)
  - Web Scraping (siehe z.B. Mancosu & Vegetti, 2020)
  - Datenspende (siehe z.B. Boeschoten et al., 2022; Breuer et al., 2022; Halavais, 2019)
- **Unique Identifier** für Verknüpfung
  - User Names auf Social-Media-Plattformen u.U. nicht unique
  - User Names können sich ändern; stabile User IDs i.d.R. nicht bekannt
  - Bei Abfrage in Surveys: Bewusste oder versehentliche Falschangaben
  - Erstellung & Nutzung eigener ID-Variablen bei Nutzung zusätzlicher Tools
- (Systematischer) **Dropout** und dadurch möglicherweise (weitere) Verzerrungen (Biases)

gesis Leibniz-Institut für Sozialwissenschaften

Dropout Stages beim Linking

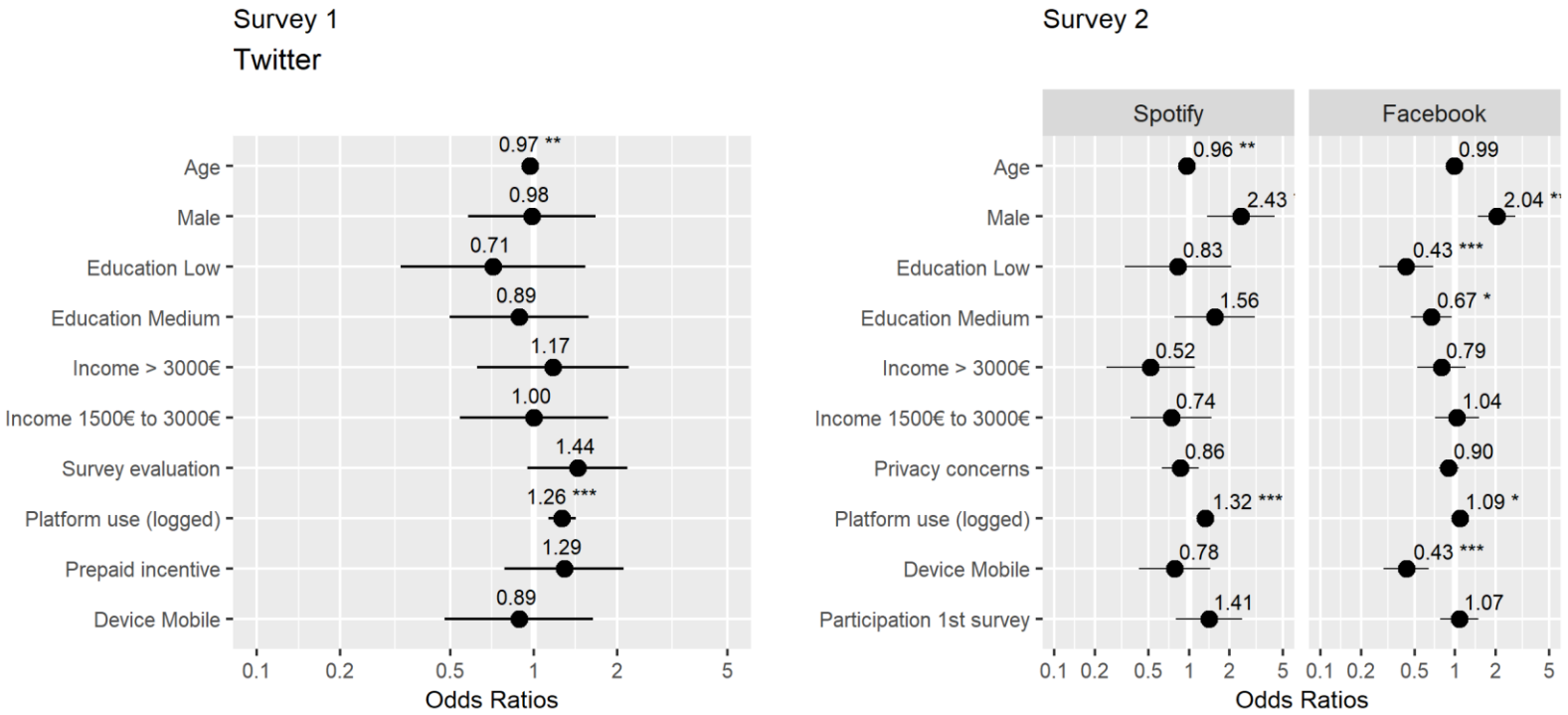


Quelle: Silber et al., 2022

Gründe für Dropout & mögliche Verzerrungen:

- Plattformnutzung
- Zustimmung (Consent)
- Technische Probleme

Bereitschaft zum Teilen von Daten



Quelle: Silber et al., 2022

- Aufwand (respondent burden) = wichtiger Faktor
- Zusätzlicher Befund aus zweiter Studie (Twitter + Health Apps): Höhe des Incentives spielt eine Rolle

Rechtliche & ethische Herausforderungen

- Erlauben **Nutzungsbedingungen (Terms of Service)** von Anwendungen/Plattformen oder APIs die Verknüpfung?
  - Bsp. Twitter Developer Policies: „We limit the circumstances under which you may match a person on Twitter to information obtained or stored off-Twitter […] You may only do this if you have express opt-in consent from the person before making the association, or as described below.“
- **Informierte Einwilligung (Informed Consent)**
  - Teilnehmer:innen müssen u.a. über Art und Zweck der Datensammlung und -nutzung informiert werden
  - Balance zwischen (notwendiger) Information und Überforderung durch Details (→ Abbruchrisiko)
- **Datenschutz**
  - Konformität mit DSGVO
  - u.U. hohes (Re-)Identifikationsrisiko bei verknüpften Daten

Lösungsansätze: Unique Identifier

- Falls möglich: (Stabile) **User IDs statt User Names** nutzen (bei vielen Plattformen wie z.B. Twitter über API abrufbar)
- Um Falschangaben zu vermeiden:
  - **Format-Checks** für Angaben in Surveys
  - **Eigenen Account für Studie erstellen** und Teilnehmer:innen bitten, diesem zu folgen oder eine Direktnachricht zu schicken

Lösungsansätze: Informed Consent

- Orientierung an **Vorlagen** (siehe z.B. Breuer et al., 2021)
- Wenn möglich: Prüfung/Beratung durch Ethikkommission
- Ggf. **Juristische Prüfung**/Rechtsberatung einholen
- Zusätzliche Detailinformationen bzw. technische Informationen auslagern und optional verfügbar machen (z.B. über Link oder Popup, der/das geöffnet werden kann)

Lösungsansätze: Datenschutz

- Ggf. **Juristische Prüfung**/Rechtsberatung einholen
- **Sicherheitsmaßnahmen** bei der Datenspeicherung: Kontrollierter Zugang, Passwörter, Verschlüsselung
- Daten so speichern und verarbeiten, dass (Re-) **Identifikationsrisiken minimiert** werden
  - Daten möglichst separat halten und nur verknüpfen, was für konkrete Analysen nötig ist

gesis Leibniz-Institut für Sozialwissenschaften

In der Speicherung getrennt, in der Analyse vereint

Szenario: Survey first, dort Abfrage von Consent und User ID für Social-Media-Plattformen oder andere Anwendungen

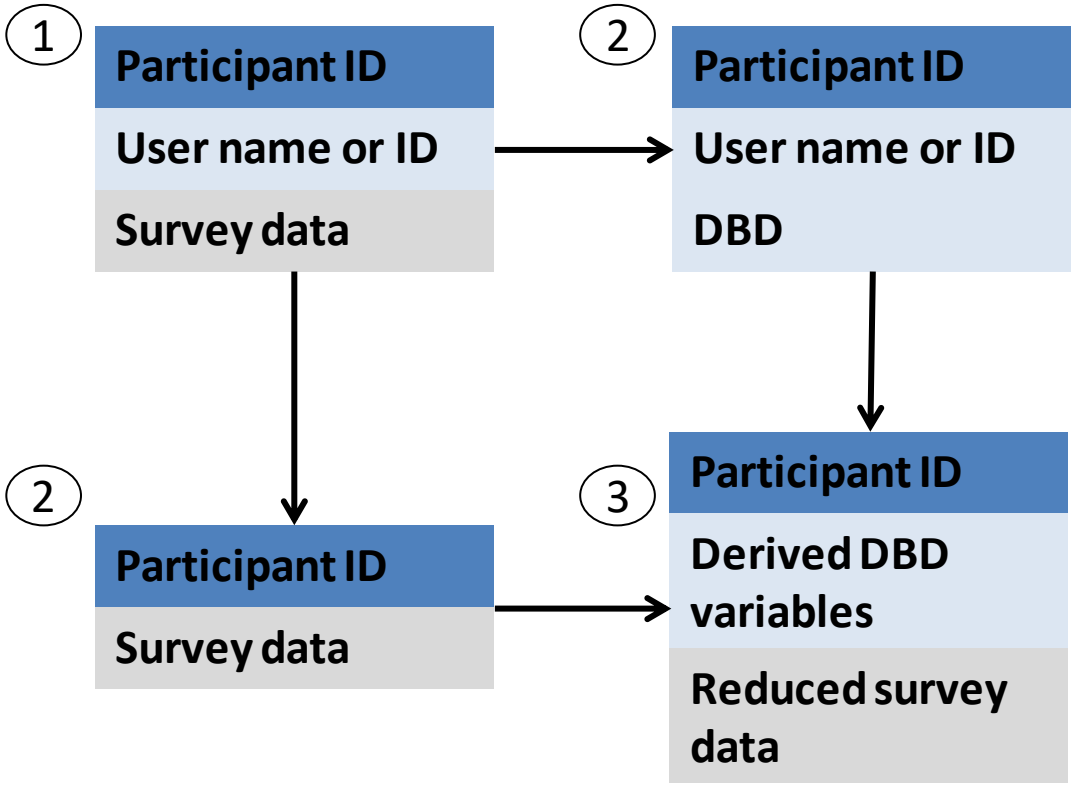


Abbildung basierend auf Beuthner et al., 2021 sowie Sloan et al., 2020

Zusammenfassung

- Die Verknüpfung von Umfragedaten und digitalen Spurdaten erlaubt es, die spezifischen Stärken dieser beiden Datentypen zu kombinieren
- Mit diesen verknüpften Daten lassen sich neue Forschungsfragen beantworten bzw. neue Antworten auf bestehende Fragen finden
- Es gibt unterschiedliche Arten der Verknüpfung: ex-ante & ex-post; probabilistisch & deterministisch; Aggregat & individuell
- Bei der Verknüpfung von Umfragen und DVD gibt es verschiedene Herausforderungen: forschungspraktische, rechtliche und ethische
- Um produktiv mit verknüpften Befragungs- und digitalen Verhaltensdaten arbeiten zu können, müssen diese Herausforderungen adäquat adressiert werden

gesis Leibniz-Institut für Sozialwissenschaften

Literatur

Beuthner, C., Breuer, J., & Jünger, S. (2021). Data Linking—Linking survey data with geospatial, social media, and sensor data. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_039

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, *4*(2), 388–423. https://doi.org/10.5117/CCR2022.2.002.BOES

Breuer, J., Al Baghal, T., Sloan, L., Bishop, L., Kondyli, D., & Linardis, A. (2021). Informed consent for linking survey and social media data—Differences between platforms and data types. *IASSIST Quarterly*, *45*(1), 1–27. https://doi.org/10.29173/iq988

Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2022). User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, Advance online publication. https://doi.org/10.1080/1369118x.2022.2097015

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

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, 1–15. https://doi.org/10.1080/1369118x.2019.1627386

Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, *6*(3). https://doi.org/10.1177/2056305120940703

Resnick, P., Adar, E., & Lampe, C. (2015). What Social Media Data We Are Missing and How to Get It. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 192–206. https://doi.org/10.1177/0002716215570006

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

Silber, H., Breuer, J., Beuthner, C., Gummer, T., Keusch, F., Siegers, P., Stier, S., & Weiß, B. (2022). Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *185*, 387–407. https://doi.org/10.1111/rssa.12954

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics*, *15*(1–2), 63–76. https://doi.org/10.1177/1556264619853447

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