Digitale Datenspuren nutzbar machen

Datenspenden als Methode der Kommunikationswissenschaft

Sitzung 2: Einführung: digitale Datenspuren

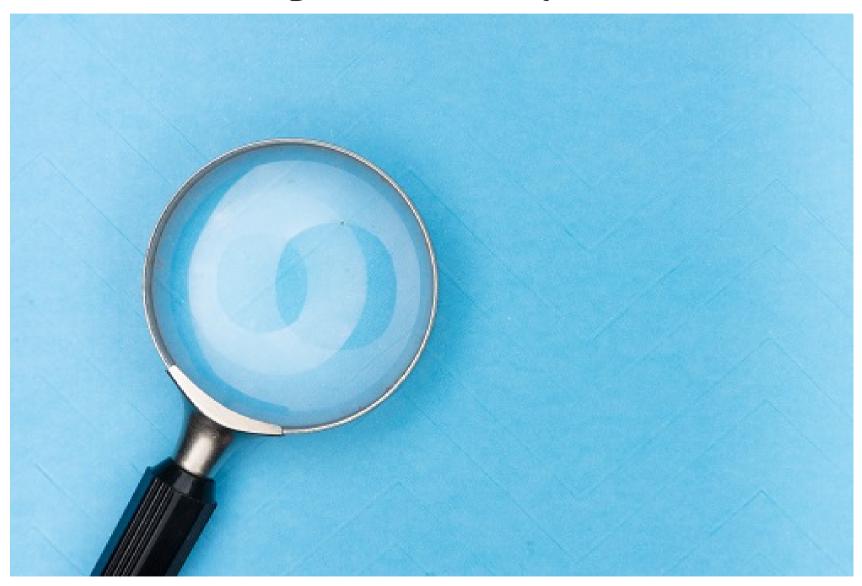
Valerie Hase (Ludwig-Maximilians-Universität München)

github.com/valeriehase & valerie-hase.com



Agenda

- 1. Was sind digitale Datenspuren?
- 2. Mit welchen Methoden kann ich auf digitale Datenspuren zugreifen?



Quelle: Foto von Markus Winkler auf Unsplash

Kennt ihr Beispiele für digitale Datenspuren? 🧽



Beispiel I

Article

Towards New Standards? Interaction Patterns of German Political Journalists in the Twittersphere

The International Journal of Press/Politics 2023, Vol. 28(1) 257–277

© The Author(s) 2021

Article reuse guidelines: sagepub.com/journals-permissions

DOI: 10.1177/19401612211025502
journals-sagepub.com/home/hij

SSAGE

Nina F. Schumacher^I, Peter Maurer^I, and Christian Nuernbergk^I

Abstract

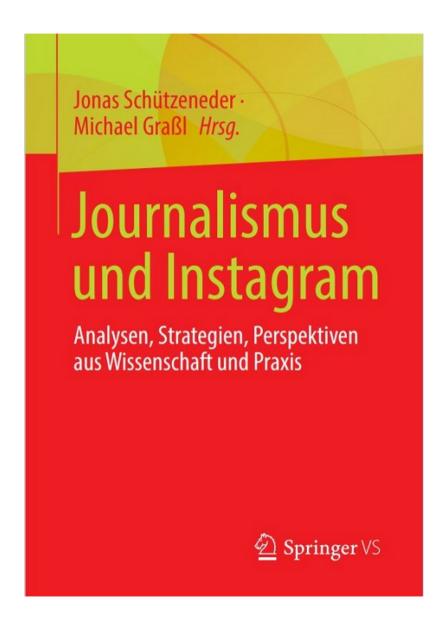
Research has provided few findings on the patterns of communication and interaction between political and media actors in social media and how these interactions have evolved. The present study analyzes how journalists registered in the German Federal Press Conference interact with members of the German Bundestag (member of parliament) on Twitter and what type of content they exchange with each other. Based on two time periods (2016, 2020), the communicative practices and the patterns of interactions are examined. New insights into the conversational structure are generated through a combination of content and network analysis. The results indicate a trend in the use of Twitter among the actors as a political opinion medium. While the dissemination of opinions among politicians remains stable, journalists use Twitter interactions more to express a subjective and critical view and less for news coverage over time. Furthermore, the analysis notes that journalists generally comment in negative terms on the parties, except for the Greens. To some extent, the results indicate that the usual journalistic norms of objectivity and balance do not apply for interactions in the Twittersphere. To evaluate whether this pattern also applies to other countries, more comparative investigations in the domain of media-politics interactions on Twitter are needed.

Keywords

political-media interaction patterns, journalistic norms, Twitter, political communication, content and network analysis



Beispiel II





Beispiel III

Original Article

Different platforms, different plots? The Kremlin-controlled search engine Yandex as a resource for Russia's informational influence in Belarus during the COVID-19 pandemic

Daria Kravets[®], Anna Ryzhova[®], Florian Toepfl and Arista Beseler[®]
University of Passau, Germany

Abstract

Extant research demonstrated that the algorithms of the Kremlin-controlled search engine Yandex, compared to those of its US-based counterpart Google, frequently produce results that are biased toward the interests of Russia's ruling elites. Prior research, however, audited Yandex's algorithms largely within Russia. In contrast, this study is the first to assess the role of Yandex's web search algorithms as a resource for Russia's informational influence abroad. To do so, we conduct a comparative algorithm audit of Google and Yandex in Belarus, examining the visibility and narratives of COVID-19-related conspiracy theories in their search results. By manually analysing the content of 1320 search results collected in mid-April to mid-May 2020, we find that, compared with Google, (1) Yandex retrieves significantly more conspiratorial content (2) that close to exclusively suspects US plotters to be behind the pandemic, even though the virus spread from the Chinese city of Wuhan across the globe.

Keyword

Search engines, conspiracy theories, Yandex, Google, Russia, foreign influence

LIMU MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Definition \mathbb{Q} : Aufzeichnung und Speicherung von Aktivitäten auf digitalen Plattformen, die Rückschlusse auf digitale wie analoge Phänomene ermöglichen

- "records of activity (trace data) undertaken through an online information system" (Howison et al., 2011, S. 2)
- "individuals leave behavioural residue (unconscious traces of actions [.]) when they interact online" (Hinds & Joinson, 2018, S. 2)

Definition 🖓 : Aufzeichnung und Speicherung von Aktivitäten auf digitalen Plattformen, die Rückschlusse auf digitale wie analoge Phänomene ermöglichen

- z.B. Tweets, Likes, Shares
- z.B. Geo-Daten (Standort teilen, Sportaktivitäten)
- z.B. digitale Zahlungen
- z.B. Spotify-Playlisten

Definition 🔐: Aufzeichnung und Speicherung von Aktivitäten auf digitalen Plattformen

• z.B. Tweets, **Likes**, Shares

Beispiel: Instagram Like



Definition 🔐: Aufzeichnung und Speicherung von Aktivitäten auf digitalen Plattformen

• z.B. Tweets, **Likes**, Shares

Beispiel: Instagram Like

```
*liked_posts - Editor
                                                                              X
Datei Bearbeiten Format Ansicht Hilfe
  "likes media likes": [
      "title": "tagesschau",
      "string_list_data": [
           "href": "https://www.instagram.com/p/Cwwp6TyIETJ",
           "value": "\u00f0\u009f\u0091\u008d",
           "timestamp": 1688963882
```

Wo lassen sich digitale Datenspuren finden?

- Social Media Plattformen (z.B. Instagram)
- Apps (z.B. Lauf-Apps)
- Payment-Systeme (z.B. Paypal)
- Wearable Devices (z.B. Smart Watch)

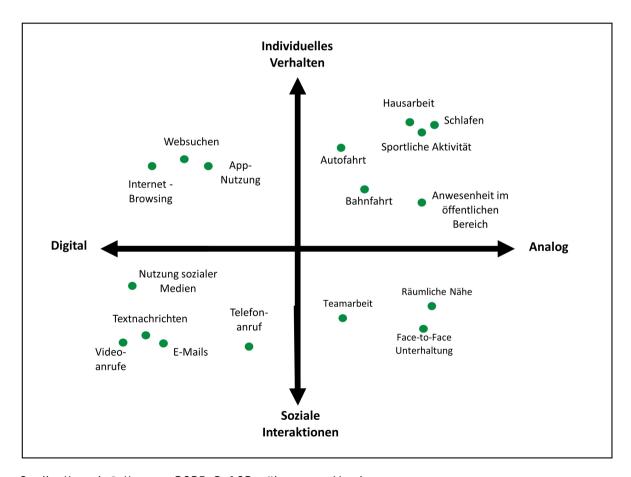
Welche Daten enthalten digitale Datenspuren?

Je nach Datenzugang u.a. (Haim & Hase, 2023; Keusch & Kreuter, 2021; Ohme et al., 2023):

- digitale Nutzerprofile/Einstellungen
- digitale Aktivitäten (Nutzung, Nachrichten, Engagement, etc.)
- digitales Targeting (Werbung, algorithmische inferierte Interessen)
- analoge Aktivitäten (Reisen, Schlafen, Sport)

Was können wir mit digitalen Datenspuren messen?

- von individuellem Verhalten zu sozialen Interaktionen
- von digitalem zu analogem Verhalten



Quelle: Keusch & Kreuter, 2023, S. 102 - übersetzte Version

Was können wir mit digitalen Datenspuren messen?

- Internet-/Smartphone-Nutzung (Ohme et al., 2021; Scharkow, 2016; Wu-Ouyang & Chan, 2022)
- Nachrichten-Nutzung (Reiss, 2022; Thorson et al., 2021)
- Prozesse öffentlicher Meinungsbildung (Jürgens & Stark, 2022; Yan et al., 2022)

- Probleme bei Selbstauskünften, z. B. bei Umfragen
- Verfügbarkeit

- Probleme bei Selbstauskünften, z. B. bei Umfragen
- Verfügbarkeit

• Probleme bei Selbstauskünften, z. B. bei Umfragen

"Wie viele Minuten am Tag nutzen Sie das Internet, um Nachrichten zu konsumieren?"



Quelle: Foto von Scott Graham auf Unsplash

- "Internet"?
- "Nachrichten"?
- "wie viele Minuten"?



- Probleme bei Selbstauskünften, z. B. bei Umfragen
 - Selbstauskünfte wenig akkurat bzw. verzerrt Datenspuren versprechen genauere Messungen (Parry et al., 2021; Scharkow, 2016; Wu-Ouyang & Chan, 2022)
 - sinkende Teilnahmebereitschaft bei Unfragen (Luiten et al., 2020)

- Probleme bei Selbstauskünften, z. B. bei Umfragen
- Verfügbarkeit
 - kostengünstig (z. B. APIs)
 - grosse Menge an Daten ("Big Data")

- Probleme bei Selbstauskünften, z. B. bei Umfragen
- Verfügbarkeit

Aber: nicht alle dieser Punkte treffen tatsächlich zu bzw. sind vorteilhaft

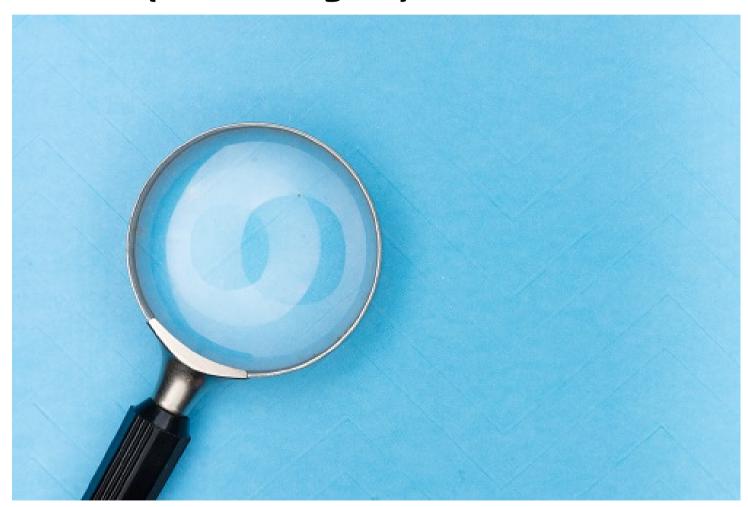
Vor- und Nachteile digitaler Datenspuren

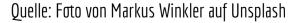
- **a**kkuratere Messungen durch Zeitstempel
- z.T. Messung neuer Variablen (z.B. zu algorithmischer Inferenz)
- **X** weiterhin Verzerrungen durch Stichproben- und Messfehler
- **X** unklare theoretische Rückbindung
- 💢 z.T. hohe Kosten für Implementierung
- mehr Daten heisst nicht bessere Daten!

Zusammenfassung: Digitale Datenspuren 🚝

- **Definition**: Aufzeichnung und Speicherung von Aktivitäten auf digitalen Plattformen (z.B. Nutzung, Engagement mit Inhalten), die Rückschlusse auf digitale wie analoge Phänomene ermöglichen
- Weiterführende Literatur:
 - Keusch & Kreuter (2021)
 - Haim & Hase (2023)
 - Ohme et al. (2023)

2. Mit welchen Methoden kann ich auf digitale Datenspuren zugreifen?







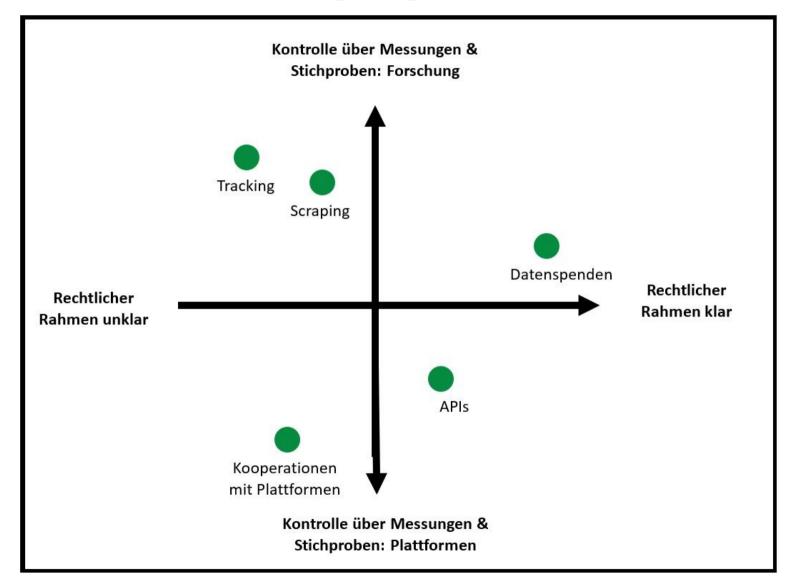
Welche methodischen Zugänge kennt ihr, um digitale Datenspuren zu sammeln?



Methodische Zugänge

- API (Jünger, 2021)
- Datenspenden (Driel et al., 2022)
- Kooperationen mit Plattformen (Wagner, 2023)
- Scraping (Mitchell, 2018)
- Tracking (Christner et al., 2022)

Methodische Zugänge





Technische Veränderungen

- Einschränkung **plattformseitiger** Zugänge zu Daten
 - Einstellung zahlreicher APIs (Bruns, 2019; Freelon, 2018)
 - Sorge über Verzerrung bei Zugang über APIs/Scraping (Buehling, 2023; Ho, 2020; Schatto-Eckrodt, 2022)
 - Kooperationen mit Plattformen sehr eingeschränkt möglich (Wagner, 2023)
- Aufkommen nutzerseitiger Zugänge
 - Datenspende
 - Tracking



Rechtliche Veränderungen

- EU verankert Recht auf eigene Daten in Art. 15 Datenschutz-Grundverordnung, kurz DSGVO
 - "Die betroffene Person hat [...] ein Recht auf Auskunft über diese personenbezogenen Daten" (Art. 15 (1)
 - "Der Verantwortliche stellt eine Kopie der personenbezogenen Daten [...] zur Verfügung" (Art. 15 (3))
- Nutzer:innen müssen laut Art. 20 **Daten übermitteln** können: "Die betroffene Person hat das Recht, die sie betreffenden personenbezogenen Daten [...] in einem strukturierten, gängigen und maschinenlesbaren Format zu erhalten" (Art. 20 (1))

Rechtliche Veränderungen

- EU verankert **Recht auf eigene Daten** in Art. 15 Datenschutz-Grundverordnung, kurz DSGVO
 - "Die betroffene Person hat [...] ein Recht auf Auskunft über diese personenbezogenen Daten" (Art. 15 (1)
 - "Der Verantwortliche stellt eine Kopie der personenbezogenen Daten [...] zur Verfügung" (Art. 15 (3))
- Nutzer:innen müssen laut Art. 20 **Daten übermitteln** können: "Die betroffene Person hat das Recht, die sie betreffenden personenbezogenen Daten [...] in einem strukturierten, gängigen und maschinenlesbaren Format zu erhalten" (Art. 20 (1))

Lösung: Plattformen bieten **Daten-Pakete** (DDPs) an, die Informationen über Nutzer:innen enthalten und von diesen heruntergeladen werden können.

Konsequenz: Die Wissenschaft nutzt diese DDPs im Rahmen von **Datenspende-Studien**.

Zusammenfassung: Datenzugänge 😂

Zusammenfassung:

- zentrale Methoden u.a. APIs, Datenspenden, Kooperationen mit Plattformen, Scraping, Tracking
- zentrale Unterschiede: Kontrolle über Stichproben & Messungen durch Plattformen, Forschung (& Nutzer:innen);
 rechtlicher Rahmen

Weiterführende Literatur:

- Haim & Hase (2023)
- Ohme et al. (2023)



Fragen?



- 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
- Buehling, K. (2023). Message Deletion on Telegram: Affected Data Types and Implications for Computational Analysis. *Communication Methods and Measures*, 1–23. https://doi.org/10.1080/19312458.2023.2183188
- Christner, C., Urman, A., Adam, S., & Maier, M. (2022). Automated Tracking Approaches for Studying Online Media Use: A Critical Review and Recommendations. *Communication Methods and Measures*, *16*(2), 79–95. https://doi.org/10.1080/19312458.2021.1907841
- Driel, I. I. van, Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures*, 1–17. https://doi.org/10.1080/19312458.2022.2109608
- Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506
- Haim, M., & Hase, V. (2023). Computational Methods und Tools für die Erhebung und Auswertung von Social-Media-Daten. In S. Stollfuß, L. Niebling, & F. Raczkowski (Eds.), *Handbuch Digitale Medien und Methoden* (pp. 1–20). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36629-2_41-1
- Hinds, J., & Joinson, A. N. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PLOS ONE*, *13*(11), e0207112. https://doi.org/10.1371/journal.pone.0207112



- Ho, J. C.-T. (2020). How biased is the sample? Reverse engineering the ranking algorithm of Facebook's Graph application programming interface. *Big Data & Society, 1*(1), 205395172090587. https://doi.org/10.1177/2053951720905874
- Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, 12(12), 767–797. https://doi.org/10.17705/1jais.00282
- Jünger, J. (2021). A brief history of APIs. In *Handbook of Computational Social Science, Volume 2* (1st ed., pp. 17–32). Routledge. https://doi.org/10.4324/9781003025245-3
- Jürgens, P., & Stark, B. (2022). Mapping Exposure Diversity: The Divergent Effects of Algorithmic Curation on News Consumption. *Journal of Communication*, jqac009. https://doi.org/10.1093/joc/jqac009
- Keusch, F., & Kreuter, F. (2021). Digital trace data. In *Handbook of Computational Social Science, Volume 1* (1st ed., pp. 100–118). Routledge. https://doi.org/10.4324/9781003024583-8
- Luiten, A., Hox, J., & Leeuw, E. de. (2020). Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics*, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025
- Mitchell, R. (2018). Web scraping with Python: Collecting more data from the modern web (Second edition). O'Reilly.
- Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking. *Communication Methods and Measures*. https://doi.org/10.1080/19312458.2023.2181319
- Ohme, J., Araujo, T., Vreese, C. H. de, & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the iOSiStreemstime flinktibin. Mobile Media & Communication, 9(2), 293–313.

- https://doi.org/10.1177/2050157920959106
- Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, *5*(11), 1535–1547. https://doi.org/10.1038/s41562-021-01117-5
- Reiss, M. V. (2022). Dissecting Non-Use of Online News Systematic Evidence from Combining Tracking and Automated Text Classification. *Digital Journalism*, 1–21. https://doi.org/10.1080/21670811.2022.2105243
- 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
- Schatto-Eckrodt, T. (2022). Hidden biases The effects of unavailable content on Twitter on sampling quality. In *Grenzen, Probleme und Lösungen bei der Stichprobenziehung* (pp. 178–195). Halem.
- Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2021). Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society, 24*(2), 183–200. https://doi.org/10.1080/1369118X.2019.1642934
- Wagner, M. W. (2023). Independence by permission. *Science*, *381*(6656), 388–391. https://doi.org/10.1126/science.adi2430
- Wu-Ouyang, B., & Chan, M. (2022). Overestimating or underestimating communication findings? Comparing self-reported with log mobile data by data donation method. *Mobile Media & Communication*, 205015792211371. https://doi.org/10.1177/20501579221137162
- Yan, P., Schroeder, R., & Stier, S. (2022). Is there a link between climate change scepticism and populism? An analysis of web tracking and survey data from Europe and the US. *Information, Communication & Society, 25*(10), 1400–1439. https://doi.org/10.10809/1999198X-2020095K-2023 Valerie Hase