Follow the User?!

Data Donation Studies for Collecting Digital Trace Data

Session 1 : Welcome & Intro to Digital Traces

Frieder Rodewald (University of Mannheim) & Valerie Hase (LMU Munich)

👉 Part of the SPP DFG Project Integrating Data Donations in Survey Infrastructure

Agenda

1. Intro to the workshop
2. What is digital trace data?
3. How can we collect digital traces?

1. Intro

Source: Image by Markus Winkler via Unsplash

Who are you?

Please raise your hand ✋ if you ....

- are familiar with the term digital trace data
- have worked with APIs
- have worked with data donation
- have worked with automated content analysis
- regularly use programming languages (e.g., R, Python)

Who are we: Frieder Rodewald

PhD, University of Mannheim (DFG project on data donation)

Research interests:

- CSS (automated content analysis, digital traces, bias)
- Privacy concerns & behavior

More info: github.com/frodew & frieder-rodewald.de

Who are we: Valerie Hase

Akademische Rätin a. Z./Postdoc, LMU Munich (prev.: University of Zurich & LSE)

Research interests:

- CSS (automated content analysis, digital traces, bias, data access)
- Digital journalism, crisis communication

More info: github.com/valeriehase & valerie-hase.com

A big thank you 🙌 to the organizers

Shoutout to the organizers behind the 7th COMPTEXT, especially

- Fabienne Lind
- Veronika Ebner
- Marcin Stecker

What is the goal of this workshop?

- ✅ Understanding digital data traces as a *type* of data
- ✅ Understanding data donation as a *method* of data access
- ✅ Working through key steps of data donation methods (user & researcher view)
- ✅ Discussing when (not) to use data donation studies
- ❌ Detailed implementation (e.g., server set-up)

Timetable

⏰ 10:00–10:20	Session 1: Welcome & Intro to Digital Traces
⏰ 10:20–11:00	Session 2: Data Donation Studies (Participant Perspective)
⏰ 11:00–12:15	Session 3: Data Donation Studies (Researcher Perspective)
⏰ 12:15–12:45	Session 4: Bias in Data Donation Studies
⏰ 12:45–13:00	Session 5: Wrap-Up

2. What is digital trace data?

Source: Image by Markus Winkler via Unsplash

Which examples for digital trace data you know? 🤔

What is digital trace data?

Definition 💡: *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*

- "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)

What is digital trace data?

Definition 💡: *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*

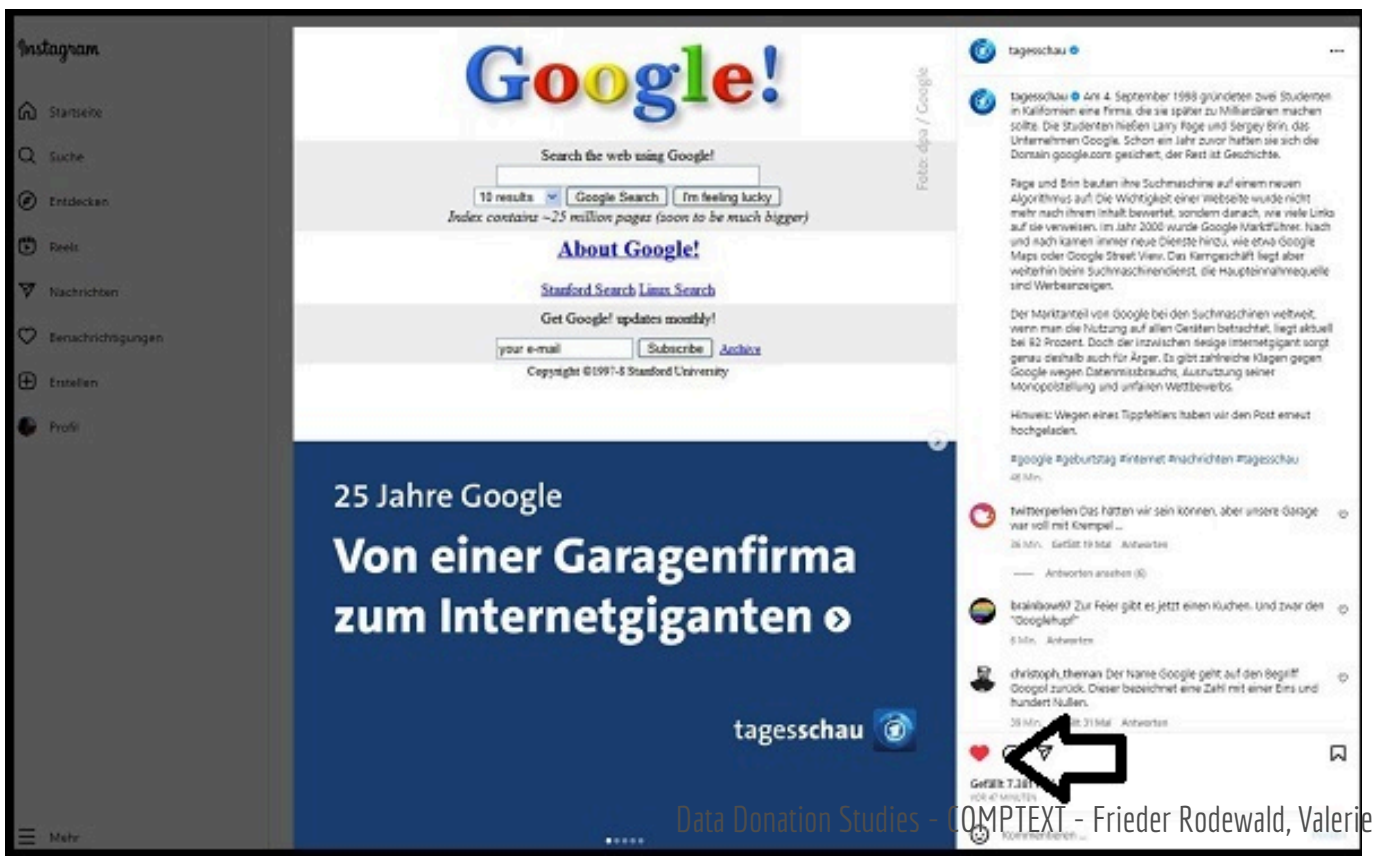
- e.g., tweets, likes, shares on social media
- e.g., geo data (locations, movements)
- e.g., digital payments
- e.g., Spotify playlists

What is digital trace data?

Definition 💡: *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*

- e.g., tweets, **likes**, shares on social media

Example: Instagram Like



What is digital trace data?

Definition 💡: *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*

- e.g., tweets, **likes**, shares on social media

Example: Instagram Like

Where can we find/collect digital trace data?

- Apps (e.g., running apps)
- Social media platforms (e.g., Instagram)
- Payment systems (e.g., Paypal)
- Wearable devices (e.g., smart watch)

Which types of data does this include?

Depending on the data collection method... (Haim & Hase, 2023; Ohme et al., 2024):

- Digital user profiles/settings (e.g., privacy settings)
- Digital activities (e.g., usage, messages, etc.)
- Digital targeting (e.g., ad exposure, algorithmically inferred interests)
- Analog activities (e.g., travelling, sleeping, sports)

Which (latent) constructs can we measure?

- Internet usage (Parry et al., 2021) related to …
  - well-being (Ohme et al., 2024)
  - voting (Bach et al., 2021)
- News usage (Reiss, 2023) related to …
  - news diversity (Jürgens & Stark, 2022)
  - public opinion formation (Yan et al., 2022)
- Movements related to …
  - Mobility during pandemics (Li et al., 2021)
  - Social networks (Sepulvado et al., 2022)

Why are digital traces becoming more popular?

- Problems with self-reported data (e.g., via survey)

“How many minutes a day do you use the internet to consume news?”

Source: Image by Scott Graham via Unsplash

- „internet”?
- „news”?
- „how many minutes”?

Why are digital traces becoming more popular?

- **Problems with self-reported data (e.g., via survey)**
  - Self-reported data may be subject to specific bias (Parry et al., 2021; Scharkow, 2016)
  - Response rates in surveys are declining (Luiten et al., 2020)

Why are digital traces becoming more popular?

- **Problems with self-reported data (e.g., via survey)**
- **Availabillity**
  - cheap (e.g., via APIs)
  - large data sets (“big data”)

Why are digital traces becoming more popular?

- Problems with self-reported data (e.g., via survey)
- Availabillity

Be careful: These "advantages" are often claimed, but **not** empirically proven.

Digital traces are neither necessarily less biased, nor cheaper, or larger (we will discuss this in Session 4 & 5).

(Dis-)advantages of digital trace data

- ✅ More fine-grained measures due to timestamps
- ✅ Partly measurement of new variables (e.g., algorithmic inference)
- ❌ Bias due to errors in representation and measurement
- ❌ Implementation can be expensive
- ❌ More data does not mean better data!

Summary: What is digital trace data? 

- **Definition**: *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*
- **Further literature**
  - Keusch & Kreuter (2021)
  - Haim & Hase (2023)
  - Ohme et al. (2024)

3. How can we collect digital traces?

Source: Image by Markus Winkler via Unsplash

Which methods do you know/have you used for collecting digital trace data*?* 🤔

Platform- and user-centric methods

- **Platform-centric** (based on platform cooperation)
  - API (Jünger, 2021)
  - Cooperation with platforms (Wagner, 2023)
- **User-centric** (based on user cooperation and informed consent)
  - Data donation (Carrière et al., 2024)
  - Linkage (Sloan et al., 2020)
  - Sensors (Struminskaya et al., 2021)
  - Tracking (Christner et al., 2022)

Platform- and user-centric methods

- Restriction of platform-centric methods
  - Discontinuation of APIs (Freelon, 2018)
  - Concern about bias (Schatto-Eckrodt, 2022; Ulloa et al., 2025)
- User-centric methods become more popular, given …
  - Changes in law (GDPR, DSA)
  - Ethical considerations (informed consent)

Changes in legal contexts

- EU secures right to own data in Art. 15 of the General Data Protection Regulation (GDPR)
  - "The data subject shall have [...] access to the personal data" (§ Art. 15, 1)
  - "The controller shall provide a copy of the personal data" (§ Art. 15, 3)
- According to § Art. 20, users must receive their data "in a structured, commonly used and machine-readable format" (§ Art. 20, 1)

Changes in legal contexts

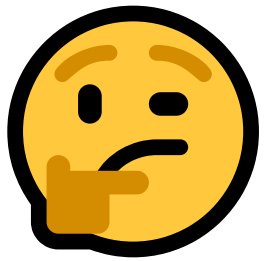
- EU secures right to own data in Art. 15 of the General Data Protection Regulation (GDPR)
  - "The data subject shall have [...] access to the personal data" (§ Art. 15, 1)
  - "The controller shall provide a copy of the personal data" (§ Art. 15, 3)
- According to § Art. 20, users must receive their data "in a structured, commonly used and machine-readable format" (§ Art. 20, 1)

👉 **Solution:** Platforms offer data download packages (DDPs), which users can request and download to inspect data.

👉 **Consequence**: Researchers uses DDPs as part of user-centric data donation studies.

Summary: How can we collect digital traces? 

- **Summary**
  - Central methods including platform-centric methods (e.g., APIs) and user-centric methods (e.g., data donation)
  - Key differences: control over samples & measurements, legal & ethical contexts
- **Further literature**
  - Haim & Hase (2023)
  - Ohme et al. (2024)

Questions? 

References

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting Voting Behavior Using Digital Trace Data. *Social Science Computer Review*, *39*(5), 862–883. https://doi.org/10.1177/0894439319882896

Carrière, T. C., Boeschoten, L., Struminskaya, B., Janssen, H. L., De Schipper, N. C., & Araujo, T. (2024). Best practices for studies using digital data donation. *Quality & Quantity*. https://doi.org/10.1007/s11135-024-01983-x

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

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://link.springer.com/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

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://www.taylorfrancis.com/books/9781003025245/chapters/10.4324/9781003025245-3

Jürgens, P., & Stark, B. (2022). Mapping Exposure Diversity: The Divergent Effects of Algorithmic Curation on News Consumption. *Journal of Communication*, *72*(3), 322–344. 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://www.taylorfrancis.com/books/9781003024583/chapters/10.4324/9781003024583-8

Li, X., Xu, H., Huang, X., Guo, C., Kang, Y., & Ye, X. (2021). Emerging geo-data sources to reveal human mobility dynamics during COVID-19 pandemic: Opportunities and challenges. *Computational Urban Science*, *1*(1), 22. https://doi.org/10.1007/s43762-021-00022-x

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

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2024). Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking. *Communication Methods and Measures*, *18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

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. (2023). Dissecting Non-Use of Online News – Systematic Evidence from Combining Tracking and Automated Text Classification. *Digital Journalism*, *11*(2), 363–383. 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.

Sepulvado, B., Wood, M. L., Fridmanski, E., Wang, C., Chandler, M. J., Lizardo, O., & Hachen, D. (2022). Predicting Homophily and Social Network Connectivity From Dyadic Behavioral Similarity Trajectory Clusters. *Social Science Computer Review*, *40*(1), 195–211. https://doi.org/10.1177/0894439320923123

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

Struminskaya, B., Lugtig, P., Toepoel, V., Schouten, B., Giesen, D., & Dolmans, R. (2021). Sharing Data Collected with Smartphone Sensors. *Public Opinion Quarterly*, *85*(S1), 423–462. https://doi.org/10.1093/poq/nfab025

Ulloa, R., Mangold, F., Schmidt, F., Gilsbach, J., & Stier, S. (2025). Beyond time delays: How web scraping distorts measures of online news consumption. *Communication Methods and Measures*, 1–22. https://doi.org/10.1080/19312458.2025.2482538

Wagner, M. W. (2023). Independence by permission. *Science*, *381*(6656), 388–391. https://doi.org/10.1126/science.adi2430

Yan, P., Schroeder, R., & Stier, S. (2022). Is there a link between climate change scepticism and populism? An