Follow the User?!

Data Donation Studies for Collecting Digital Trace Data

Session 1️⃣ : 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?

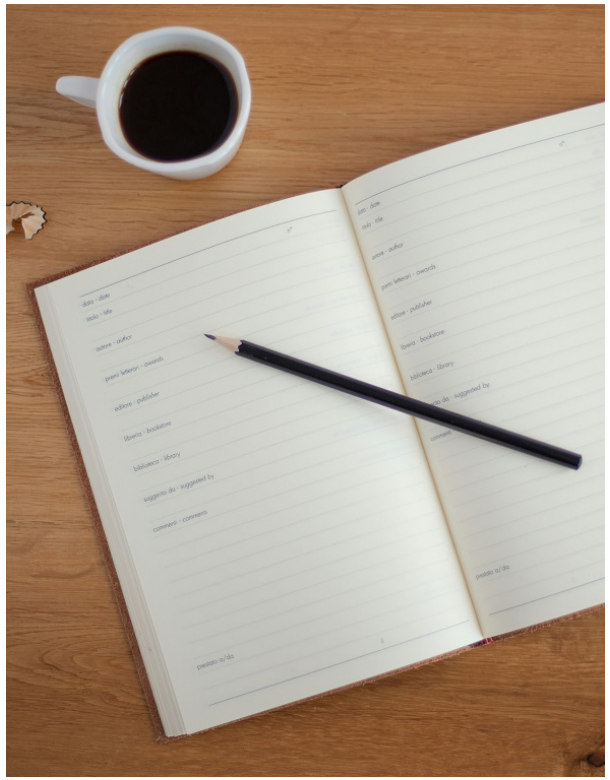


Image by Hope House Press via Unsplash

Before we start: Have you requested and downloaded your DDP? 🤔

Otherwise, please check your email and use this link: https://next.eyra.co/assignment/334/participate?participant=XXX

1. Intro (Valerie)

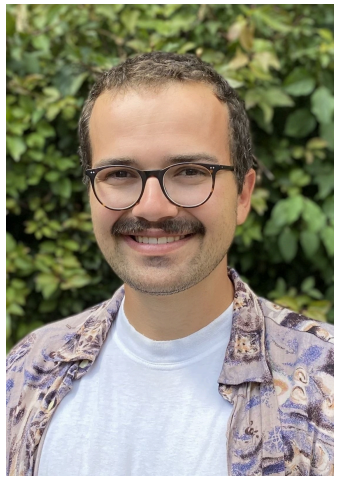
Source: Image by Markus Winkler via Unsplash

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)

About us: Frieder Rodewald



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

Research interests:

- CSS (automated content analysis, digital traces, bias)
- Privacy concerns & behavior
  "I study what people do online"

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

About us: 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, coding data extraction scripts)

Timetable

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

2. What is digital trace data? (Frieder)

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*

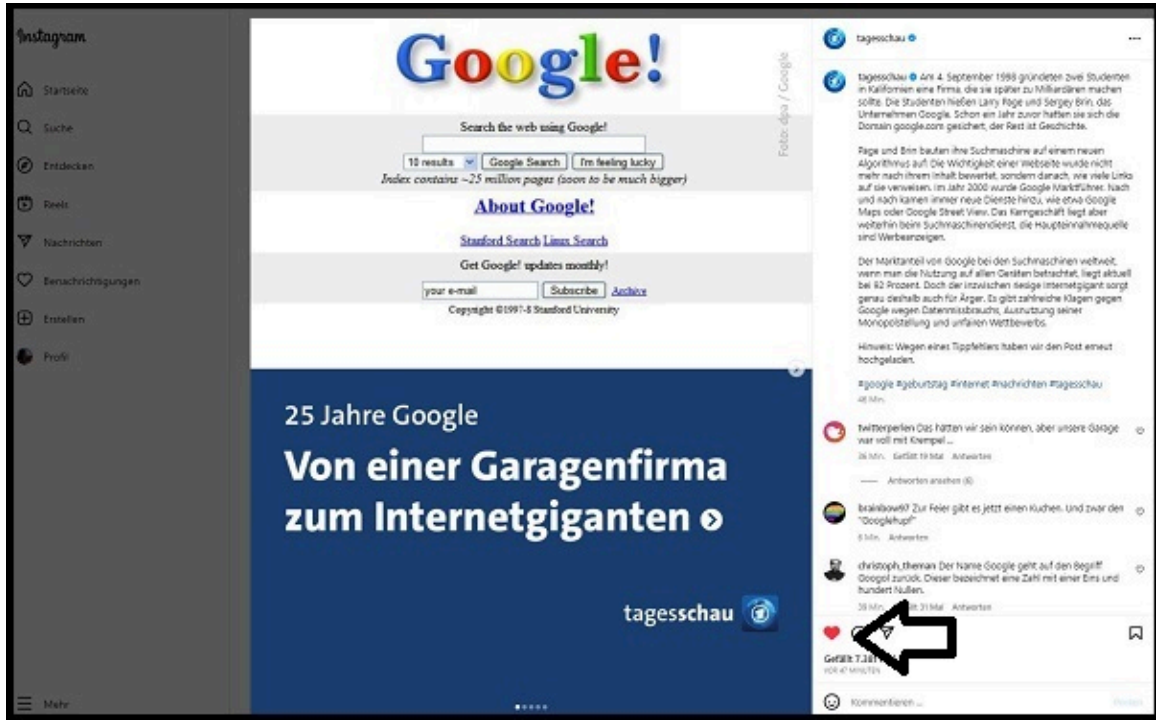
- e.g., tweets, likes, shares on social media
- e.g., geo data (locations, movements)
- e.g., digital payments
- e.g., Spotify playlists

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

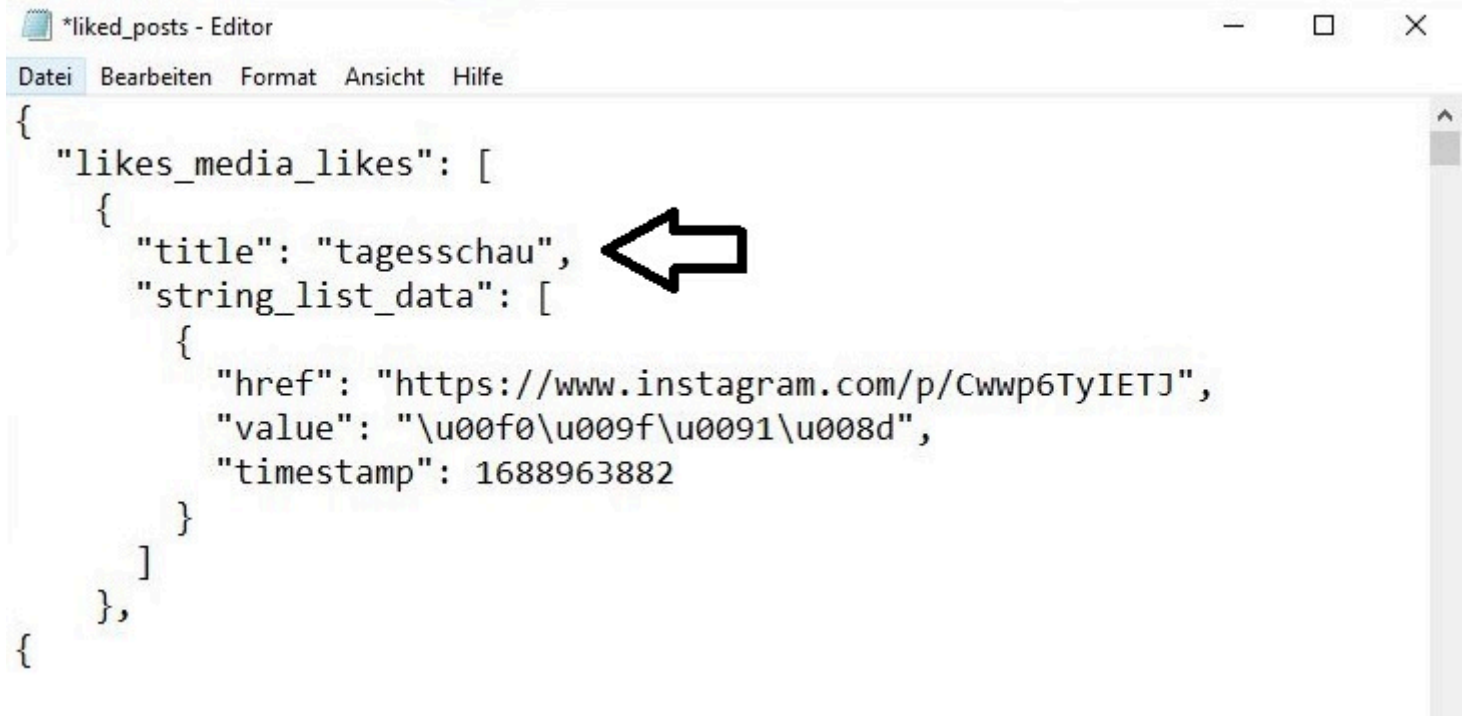


What is digital trace data?

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):

- often fine-grained (e.g., time-stamped)
- often longitudinal (e.g., over years, within-individual change)
- often less reactive (e.g., less concerns about social desirability)

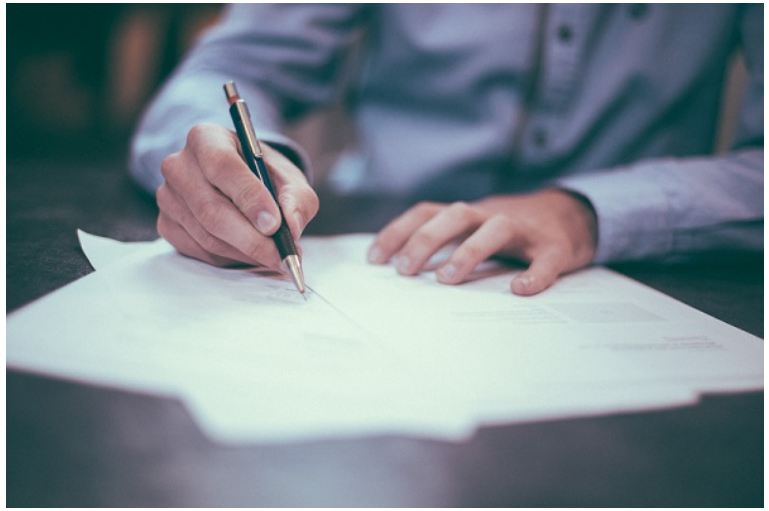
Which (latent) constructs can we measure?

- **Internet use** (Parry et al., 2021) related to ...
  - well-being (Ohme et al., 2024)
  - voting (Bach et al., 2021)
- **News engagement** (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 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 4️⃣).

(Dis-)advantages of digital trace data

- ✅ More fine-grained, often longitudinal 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 and cumbersome
- ❌ 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? (Valerie)

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) or "follow the user" approaches (Caliandro, 2024)
  - 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

- Restrictions of platform-centric methods
  - Discontinuation of APIs (Freelon, 2018)
  - Concerns about bias (Schatto-Eckrodt, 2022; Ulloa et al., 2025)
- User-centric methods become more popular, given ...
  - Changes in law that enable such studies (GDPR, DSA)
  - Presumably (!) less biased data
  - Ethical considerations (informed consent)

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

Caliandro, A. (2024). Follow the user: Taking advantage of Internet users as methodological resources. *Convergence: The International Journal of Research into New Media Technologies*, 13548565241307569. https://doi.org/10.1177/13548565241307569

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

Jünger, J. (2021). A brief history of APIs. In *Handbook of Computational Social Science, Volume 2* (1st ed., pp. 17–32).