What do people do online? Using data donation to understand digital behavior.

a workshop at the SPP Junior Researcher Meeting

Frieder Rodewald ¹⁰

University of Mannheim & Institute for Employment Research

Sebastian Prechsl 🗓

Institute for Employment Research & LMU Munich

October 22, 2025

What is data donation?

The researcher's perspective.

What are methodological decisions researchers have to take in data donation studies?

Key decisions:

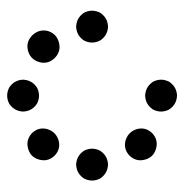
- Which theoretical questions do I want to answer?
- How do I operationalize key variables via my data donation tool?
- How do I integrate the tool in surveys & recruit participants?



Survey



Request & Download Data



Extract Data



Inspect Data



Consent

Decisions we took

Frame and Motivate:

Testing strategies to increase participation in data donation studies.

Frieder Rodewald

University of Mannheim f.rodewald@uni-mannheim.de https://orcid.org/0009-0009-6859-5761

Florian Keusch

University of Mannheim https://orcid.org/0000-0003-1002-4092

Valerie Hase

Ludwig-Maximilian-University Munich https://orcid.org/0000-0001-6656-4894

Sebastian Prechs

nstitute for Employment Research, Germany Ludwig-Maximilian-University Munich https://orcid.org/0000-0001-9033-7317

Frauke Kreuter

Ludwig-Maximilian-University Munich https://orcid.org/0000-0002-7339-2645

Mark Trappmann

Institute for Employment Research, Germany

Preprint Version

ersion: 09. October 202

- **Goal**: Increase participation in the context of labour market studies and understand non-response bias
- **Issue**: Low response rates [e.g., (Hase and Haim 2024; Keusch et al. 2024)]
 - Behavioral intentions as "willingness to donate"
 high (79-52% of survey respondents)
 - Actual behavior as "participation in data donation"
 low (37-12% of survey respondents)
 - Well known intention-behavior gap; where seems to help (Kmetty et al. 2025)
- Sample: A non-probability panel (online access panel)

B Survey

• Survey on platform usage (YouTube, Instagram, and LinkedIn), labor market characteristics and common indicators for non-participation

Data request & download

: Data extraction

X Strategy to make the extraction work

- 1. Take a look at the DDP; download it, best for multiple time periods and for different languages
- 2. Understand the structure of the JSON or CSV.
- 3. Get an example file running.
- 4. Write the code for the extraction script.
- 5. Test your script, first locally and then in the wild.
- 6. Adapt your script.

Task: Try it yourself.

Take a look at your downloaded data. What do you see; anything caught your eye?

Feel free to work in groups of 2-3 people for 5 minutes.

Different degrees in standardization for DDP content (Hase et al. 2024)...

- Documentation
 - DDP structure?
 - Measurements?
- Completeness & scope
 - Missing data?
 - Limited time frames?
 - Language sensitive?



Key issues (Hase et al. 2024)

- Missing documentation by platforms (e.g., file structure)
- Sudden changes in DDPs
- Differences across languages & devices
- Insufficient in-tool classification (e.g., LLM integration)

Example: Extract list of subscriptions

4	А	В	С	D
1	Channel Id	Channel Url	Channel Title	
2	UC0vBXGSyV14uvJ4hECDOl0Q	http://www.youtube.com/channel/UC0vBXGSyV14uvJ4hECDOl0Q	Techquickie	
3	UC1H1NWNTG2Xi3pt85ykVSHA	http://www.youtube.com/channel/UC1H1NWNTG2Xi3pt85ykVSHA	Jordan Harrod	
4	UC4NNPgQ9sOkBjw6GlkgCylg	http://www.youtube.com/channel/UC4NNPgQ9sOkBjw6GlkgCylg	Ben Vallack	
5	UC6-ymYjG0SU0jUWnWh9ZzEQ	http://www.youtube.com/channel/UC6-ymYjG0SU0jUWnWh9ZzEQ	Wisecrack	
6	UC6DUUo63tKyr1_BHN26OiJw	http://www.youtube.com/channel/UC6DUUo63tKyr1_BHN26OiJw	Wahre Verbrechen. Wahre Stories	
7	UCAD-xOOaUI6N7Uq9laOVbcw	http://www.youtube.com/channel/UCAD-xOOaUI6N7Uq9laOVbcw	Code Therapy w/ René Rebe	
8	UCAXCI-ASTfZqfv9-YklfPlA	http://www.youtube.com/channel/UCAXCI-ASTfZqfv9-YklfPIA	PacKMeN	
9	UCApPPpJ4d3ueW38lArwiWoA	http://www.youtube.com/channel/UCApPPpJ4d3ueW38lArwiWoA	Kenny Beats	
10	UCBa659QWEk1AI4TgmrJ2A	http://www.youtube.com/channel/UCBa659QWEk1AI4TgmrJ2A	Tom Scott	
11	UCDhu1klCDnf2glev0YbYkDA	http://www.youtube.com/channel/UCDhu1klCDnf2glev0YbYkDA	BeHaind	
12	UCFZms3ivokCP_HO8o5JzxEw	http://www.youtube.com/channel/UCFZms3ivokCP_HO8o5JzxEw	moTricksTV	
13	UCGII8SK7YD2B0Gd43DZk4NQ	http://www.youtube.com/channel/UCGII8SK7YD2B0Gd43DZk4NQ	mattes	
14	UCHnyfMqiRRG1u-2MsSQLbXA	http://www.youtube.com/channel/UCHnyfMqiRRG1u-2MsSQLbXA	Veritasium	
15	UCJXa3_WNNmIpewOtCHf3B0g	http://www.youtube.com/channel/UCJXa3_WNNmIpewOtCHf3B0g	LaurieWired	
16	UCJkMlOu7faDgqh4PfzbpLdg	http://www.youtube.com/channel/UCJkMlOu7faDgqh4PfzbpLdg	Nerdwriter1	
17	UCMELMEuQqmxTqM4_ArhHPjQ	http://www.youtube.com/channel/UCMELMEuQqmxTqM4_ArhHPjQ	High5	
18	UCMI9UhY1ehLGfOP5KNIKIaQ	http://www.youtube.com/channel/UCMI9UhY1ehLGfOP5KNIKIaQ	Doktor Allwissend	
19	UCMu5gPmKp5av0QCAajKTMhw	http://www.youtube.com/channel/UCMu5gPmKp5av0QCAajKTMhw	ERB	
20	UCN29LJGZ8FY30ysxdTnDsaw	http://www.youtube.com/channel/UCN29LJGZ8FY30ysxdTnDsaw	Filmanalyse	
21	UCNTwGcSEDHIbGhk7l5xFGwA	http://www.youtube.com/channel/UCNTwGcSEDHIbGhk7l5xFGwA	tinseltown	
22	UCOpcACMWblDls9Z6GERVi1A	http://www.youtube.com/channel/UCOpcACMWblDls9Z6GERVi1A	Screen Junkies	
23	UCU98JVxJf-VQXbPQPNbkbQQ	http://www.youtube.com/channel/UCU98JVxJf-VQXbPQPNbkbQQ	Meditations for the anxious mind	
24	UCUyeluBRhGPCW4rPe_UvBZQ	http://www.youtube.com/channel/UCUyeluBRhGPCW4rPe_UvBZQ	ThePrimeTime	

subscriptions.csv (before processing)

```
"subscriptions": {
    "extraction_function": ef.extract_subscriptions,
    "possible_filenames": ["Abos.csv", "subscriptions.csv"],
    "title": {
        "en": "Which channels are you subscribed to?",
        "de": "Welche Kanäle haben Sie abonniert?",
        "nl": "Op welke kanalen ben je geabonneerd?",
    },
    }
}
```

```
def extract_youtube_content_from_zip_folder(zip_file_path, possible_filenames):
    """Extract content from YouTube data export zip file using filenames"""
    try:
        with zipfile.ZipFile(zip_file_path, "r") as zip_ref:
            # Get the list of file names in the zip file
            filenames = zip_ref.namelist()
            # Look for matching files
            for possible_filename in possible_filenames:
                for filename in filenames:
                    if possible_filename in filename:
                        try:
                            # Process based on file extension
                            if filename.endswith(".json"):
                                with zip_ref.open(filename) as json_file:
                                    json_content = json.loads(json_file.read())
                                    return json_content
                            elif file_name.endswith(".csv"):
                                with zip_ref.open(file_name) as csv_file:
                                    csv_content = pd.read_csv(csv_file)
```

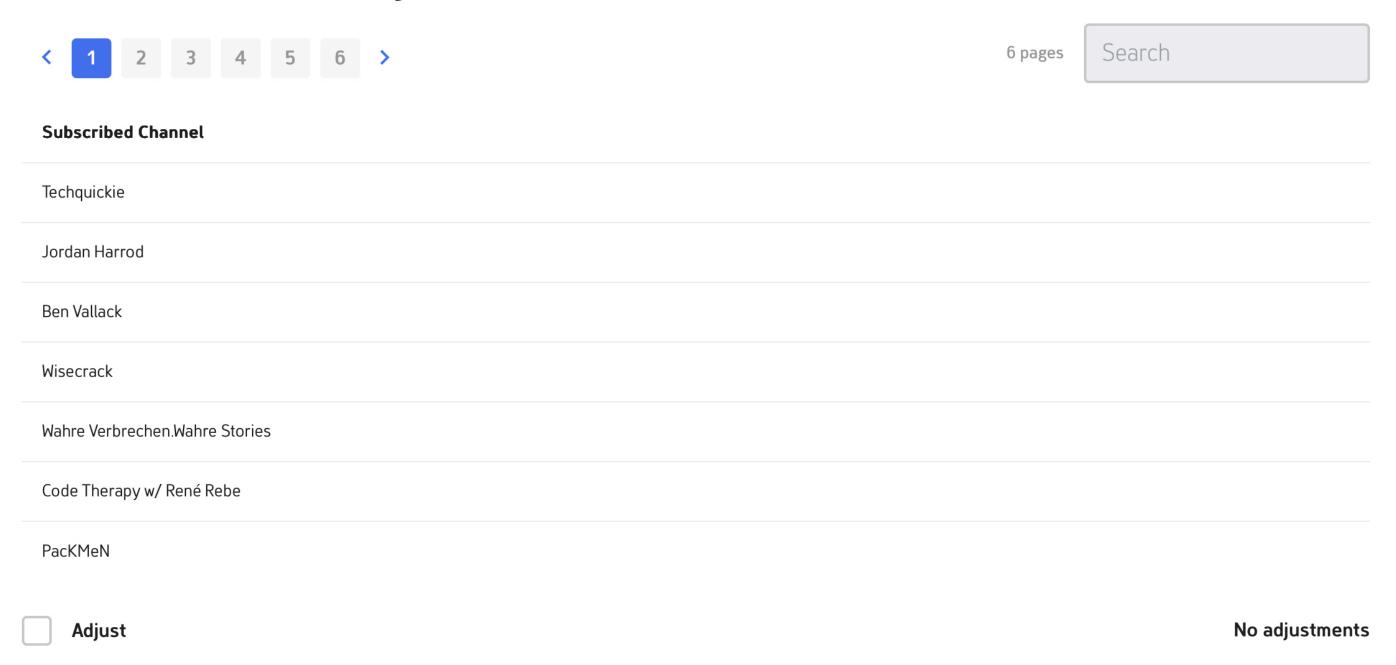
```
def extract_subscriptions(subscriptions_csv):
    """Extract YouTube channel subscriptions"""

    # Define column name
    if "Kanaltitel" in subscriptions_csv.columns:
        channel_column = "Kanaltitel"
    else:
        channel_column = "Channel Title"

# Define description
    channel_name = "Subscribed Channel"

# Create DataFrame with just the channel names
    subscriptions_df = pd.DataFrame({channel_name: subscriptions_csv[channel_column]})
    return subscriptions_df
```

Which channels are you subscribed to?



subscriptions.csv (after processing)

Q Data inspection

Data donation

- A data donation platform helps to guide them through the process
- The process should be made as easy as possible for participants



Workshop Takeaways

- There are many ways in which researchers can learn about people's online behavior through digital trace data
- People can request, download and finally donate their (anonymized) data form most online platforms
- Data donation can be burdensome for participants
- Data quailty heavily depends on the platform and what kind of data you extract from participants

Thank you for participating; happy to talk with you about data donation (and anything else) throughout the next days.

EXTRA: Can I extend the data?

- Manual annotation by participants during data donation
- APIs/scraping to extend collected data
- Text-as-data methods for classification

EXTRA: Errors in representation

For example ...

- Coverage error: Who is (not) represented in the sampling frame? (e.g., social media users vs. YouTube users)
- **Sampling error**: Who is (not) represented in the sample? (e.g., non-probability samples)
- **Non-response error**: Who does (not) want to participate in the data donation?
- **Compliance error**: Who is (not) able to participate in the data donation?

What do you think: Which participant characteristics may correlate with non-response or non-compliance?

EXTRA: What's next for data donation studies?

Advancing the method

- Multimodal & cross-platform data (Wedel, Ohme, and Araujo 2024)
- In-tool, local classification (e.g., local ML/LLMs?)
- Workflow/UX-perspective

Data as a political tool

- Platforms do (willingly?) not provide data according to the GDPR/DSA (Hase et al. 2024)
- The EU has started to sanction platforms like X/TikTok
- DSA may become the subject of larger geo-political debates with the USA (Seiling, Ohme, and De Vreese 2025)



Can we improve & apply the method?

- Can the method actually be applied for empirical research? (few examples, like (Thorson et al. 2021; Wojcieszak et al. 2024))
- Requires interdisciplinary perspectives (e.g., addressing bias, integration in probability-based panels)

Questions?

References

- Hase, Valerie, Jef Ausloos, Laura Boeschoten, Nico Pfiffner, Heleen Janssen, Theo Araujo, Thijs Carrière, et al. 2024. "Fulfilling Data Access Obligations: How Could (and Should) Platforms Facilitate Data Donation Studies?" *Internet Policy Review* 13 (3). https://doi.org/10.14763/2024.3.1793.
- Hase, Valerie, and Mario Haim. 2024. "Can We Get Rid of Bias? Mitigating Systematic Error in Data Donation Studies Through Survey Design Strategies." *Computational Communication Research* 6 (2): 1. https://doi.org/10.5117/CCR2024.2.2.HASE.
- Keusch, Florian, Paulina K. Pankowska, Alexandru Cernat, and Ruben L. Bach. 2024. "Do You Have Two Minutes to Talk about Your Data? Willingness to Participate and Nonparticipation Bias in Facebook Data Donation." *Field Methods* 36 (4): 279–93. https://doi.org/10.1177/1525822X231225907.
- Kmetty, Zoltán, Ádám Stefkovics, Júlia Számely, Dongning Deng, Anikó Kellner, Edit Pauló, Elisa Omodei, and Júlia Koltai. 2025. "Determinants of Willingness to Donate Data from Social Media Platforms." *Information, Communication & Society* 28 (7): 1324–49. https://doi.org/10.1080/1369118X.2024.2340995.
- Seiling, Lukas, Jakob Ohme, and Claes De Vreese. 2025. "Wird Europa Den DSA in Verhandlungen Mit Trump Opfern?" *Tagesspiegel*, March.
- Thorson, Kjerstin, Kelley Cotter, Mel Medeiros, and Chankyung Pak. 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.
- Wedel, Lion, Jakob Ohme, and Theo Araujo. 2024. "Augmenting Data Download Packages Integrating Data Donations, Video Metadata, and the Multimodal Nature of Audio-visual Content." *Methods, Data, Analyses (Online First)*, October, 32 Pages. https://doi.org/10.12758/MDA.2024.08.
- Wojcieszak, Magdalena, Ericka Menchen-Trevino, Bernhard Clemm Von Hohenberg, Sjifra De Leeuw, João Gonçalves, Sam Davidson, and Alexandre Gonçalves. 2024. "Non-News Websites Expose People to More Political Content Than News Websites: Evidence from Browsing Data in Three Countries." *Political*