Introduction to Data Donation

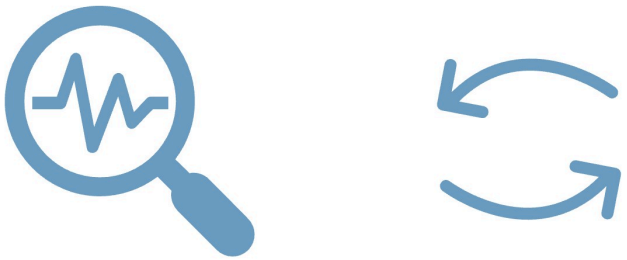
Workshop TU Ilmenau 2026

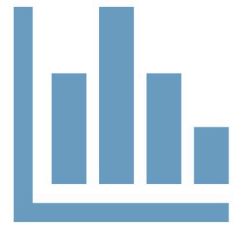
Session 3️⃣ : Data Donation Studies (Researcher Perspective)

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

Please come up with 2-3 research questions/hypotheses you may want to answer using data donation. To answer these, which methodological decisions would you have to take? 🤔

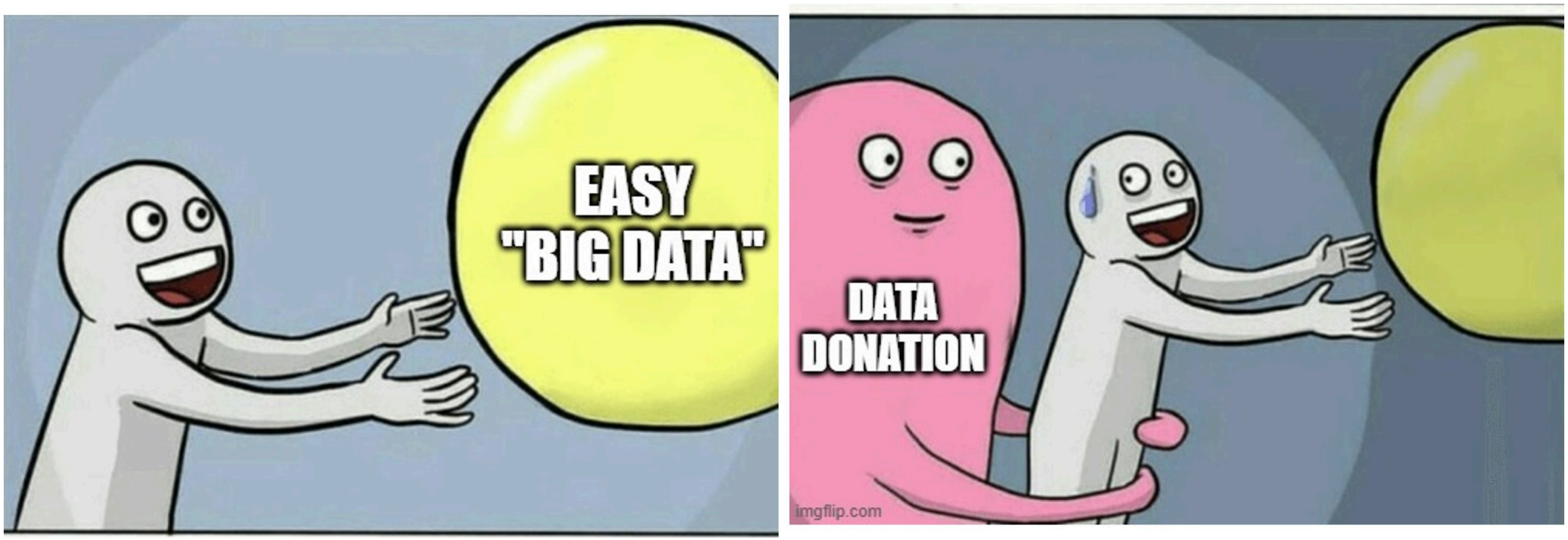
Data Donation: Methodological Decisions



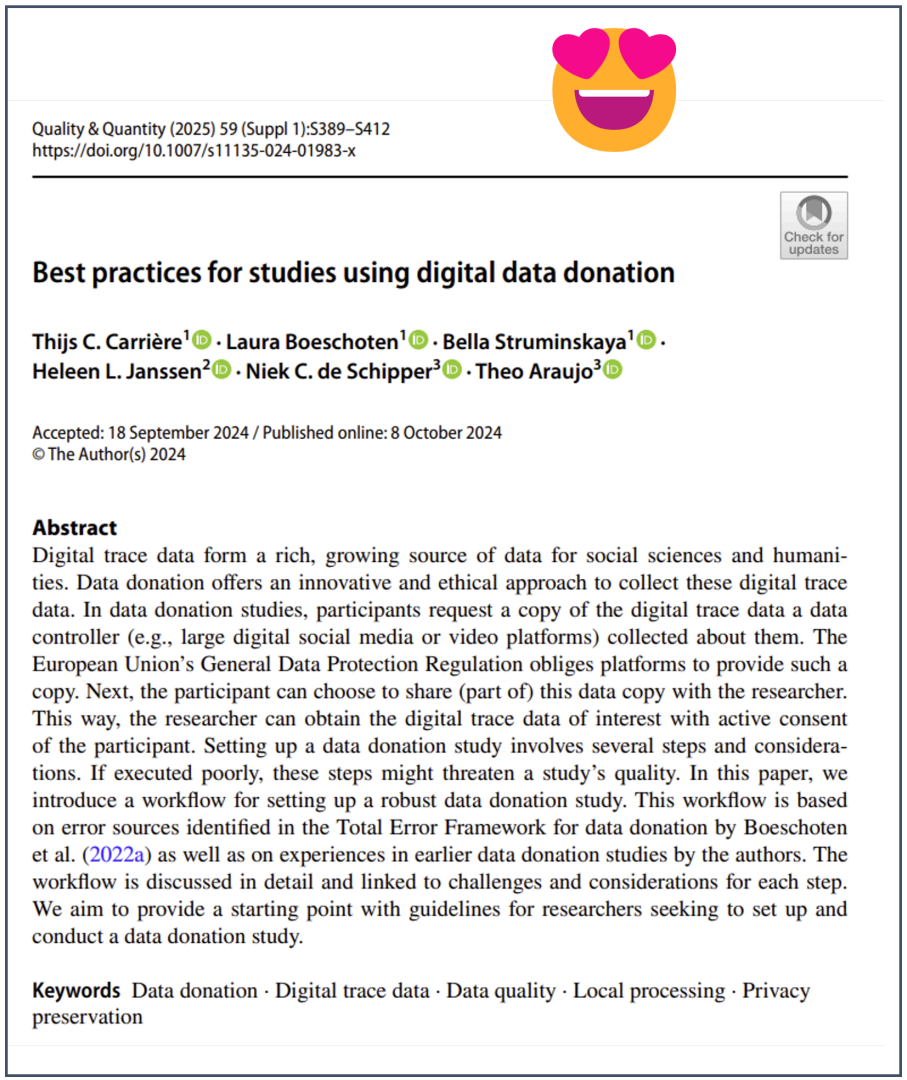


1. Research Design & Tool Set-Up
2. Data Cleaning & Augmentation
3. Modelling

Data Donation: Methodological Decisions



For a summary: shoutout to this primer



Quality & Quantity (2025) 59 (Suppl 1):S389–S412
https://doi.org/10.1007/s11135-024-01983-x

Check for updates

Best practices for studies using digital data donation

Thijs C. Carrière[1] · Laura Boeschoten[1] · Bella Struminskaya[1] · Heleen L. Janssen[2] · Niek C. de Schipper[3] · Theo Araujo[3]

Accepted: 18 September 2024 / Published online: 8 October 2024

Abstract
Digital trace data form a rich, growing source of data for social sciences and humanities. Data donation offers an innovative and ethical approach to collect these digital trace data. In data donation studies, participants request a copy of the digital trace data a data controller (e.g., large digital social media or video platforms) collected about them. The European Union's General Data Protection Regulation obliges platforms to provide such a copy. Next, the participant can choose to share (part of) this data copy with the researcher. This way, the researcher can obtain the digital trace data of interest with active consent of the participant. Setting up a data donation study involves several steps and considerations. If executed poorly, these steps might threaten a study's quality. In this paper, we introduce a workflow for setting up a robust data donation study. This workflow is based on error sources identified in the Total Error Framework for data donation by Boeschoten et al. (2022a) as well as on experiences in earlier data donation studies by the authors. The workflow is discussed in detail and linked to challenges and considerations for each step. We aim to provide a starting point with guidelines for researchers seeking to set up and conduct a data donation study.

Keywords Data donation · Digital trace data · Data quality · Local processing · Privacy preservation

(Carrière et al., 2024)

Agenda

1. **Research design & tool set-up**
2. **Data cleaning & augmentation**
3. **Modelling**
    - 📢 **Task 3**: Example Analysis of YouTube Watch history

Image by Hope House Press via Unsplash

1) Research design & tool set-up

Source: Image by Markus Winkler via Unsplash

Step I: Research design & tool set-up

1 Research Design & Tool Set-Up

2 Data Cleaning & Augmentation

3 Modelling

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables via my data donation tool?

1.3 How do I integrate the tool in surveys & recruit participants?

Step I: Research design & tool set-up

1 Research Design & Tool Set-Up

2 Data Cleaning & Augmentation

3 Modelling

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables via my data donation tool?

1.3 How do I integrate the tool in surveys & recruit participants?

Step 1.1 Which questions do I want to answer?

Step 1.1 Which questions do I want to answer?

Article

Avenues to News and Diverse News Exposure Online: Comparing Direct Navigation, Social Media, News Aggregators, Search Queries, and Article Hyperlinks

The International Journal of Press/Politics
2022, Vol. 27(4) 860–886
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/19401612211009160
journals.sagepub.com/home/hij
SAGE

Magdalena Wojcieszak[1,2], Ericka Menchen-Trevino[3], Joao F. F. Goncalves[4], and Brian Weeks[5]

Original Article

What is news? Mapping the diversity of news experiences in digital trace data

Journalism
2024, Vol. 0(0) 1–19
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14648849241303115
journals.sagepub.com/home/jou
Sage

Felicia Loecherbach
University of Amsterdam, The Netherlands

Judith Moeller
Hans-Bredow-Institut für Medienforschung an der Universität Hamburg, Hamburg, Germany

Damian Trilling
Vrije Universiteit Amsterdam, The Netherlands

Wouter van Atteveldt
Vrije Universiteit Amsterdam, The Netherlands

Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.1 (2023) 1–32
HTTPS://DOI.ORG/10.5117/CCR2023.1.15.GONG

Media selection is highly predictable, in principle

Xuanjun Gong
Department of Communication, University of California, Davis
Department of Statistics, University of California, Davis

Richard Huskey
Department of Communication, University of California, Davis
Cognitive Science Program, University of California Davis
Center for Mind and Brain, University of California Davis

Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

Do New Romantic Couples Use More Similar Language Over Time? Evidence from Intensive Longitudinal Text Messages

Miriam Brinberg[1] & Nilam Ram[2]

cogitatio

Media and Communication (ISSN: 2183–2439)
2023, Volume 11, Issue 1, Pages 19–30
https://doi.org/10.17645/mac.v11i1.6030

Article

Googling Referendum Campaigns: Analyzing Online Search Patterns Regarding Swiss Direct-Democratic Votes

Sina Blassnig *, Eliza Mitova, Nico Pfiffner, and Michael V. Reiss

Special Issue: Data Reflectivity: New Pathways in Bridging Datafication and User Studies

CONVERGENCE

Integrating trace data into interviews: Better interviews, better data

Convergence: The International Journal of Research into New Media Technologies
2024, Vol. 30(6) 2059–2074
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/13548565241300897
journals.sagepub.com/home/con
Sage

Ri Pierce-Grove
Columbia University, USA

Elizabeth Anne Watkins
Intel Labs, USA

Step I.I Which questions do I want to answer?

Empirical studies on data donation focus on ...

- exposure/pathways to news or political content (Loecherbach et al., 2024; Wojcieszak et al., 2023; Xuanjun Gong & Richard Huskey, 2023)
- social interaction/relationship development (Brinberg & Ram, 2021; Corten et al., 2025; Virtanen et al., 2021)

Step I.I Which questions do I want to answer?

Research designs include ...

- use of **observational data to describe/explain** (e.g., how platforms shape exposure diversity, often via sequential modeling) (Loecherbach et al., 2024)
- **combination with experimental designs** (e.g., interventions, sock puppet training) (Yu et al., 2024)
- **triangulation with qualitative methods** (e.g., walk-through-interviews) (Pierce-Grove & Watkins, 2024)

Step I.I Which questions do I want to answer?

Research designs include ...

- use of **observational data to describe/explain** (e.g., how platforms shape exposure diversity, often via sequential modeling) (Loecherbach et al., 2024)
- **combination with experimental designs** (e.g., interventions, sock puppet training) (Yu et al., 2024)
- **triangulation with qualitative methods** (e.g., walk-through-interviews) (Pierce-Grove & Watkins, 2024)

⚠️ Causal inference remains a key problem!

⚠️ Match between theoretical concepts and measurements remains a key problem!

Step I: Research design & tool set-up

1 Research Design & Tool Set-Up

2 Data Cleaning & Augmentation

3 Modelling

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables?

1.3 How do I integrate the tool in surveys & recruit participants?

Step I.II: How do I operationalize key variables?

Key questions:

- Which data donation tool do I use? 🖥️
- Which variables do I extract? 🔎
- How do I anonymize data? 🙈

Step I.II: Which data donation tool do I use? 🖥️

Key questions:

- **Which data donation tool do I use?** 🖥️
- Which variables do I extract? 🔎
- How do I anonymize data? 🙈

Step I.II: Which data donation tool do I use? 

- Participants “upload” data (nothing is send anywhere)
- Local extraction, anonymization, & aggregation
- Users can delete data
- Informed consent, only then: send to researcher server

Step I.II: Which data donation tool do I use? 

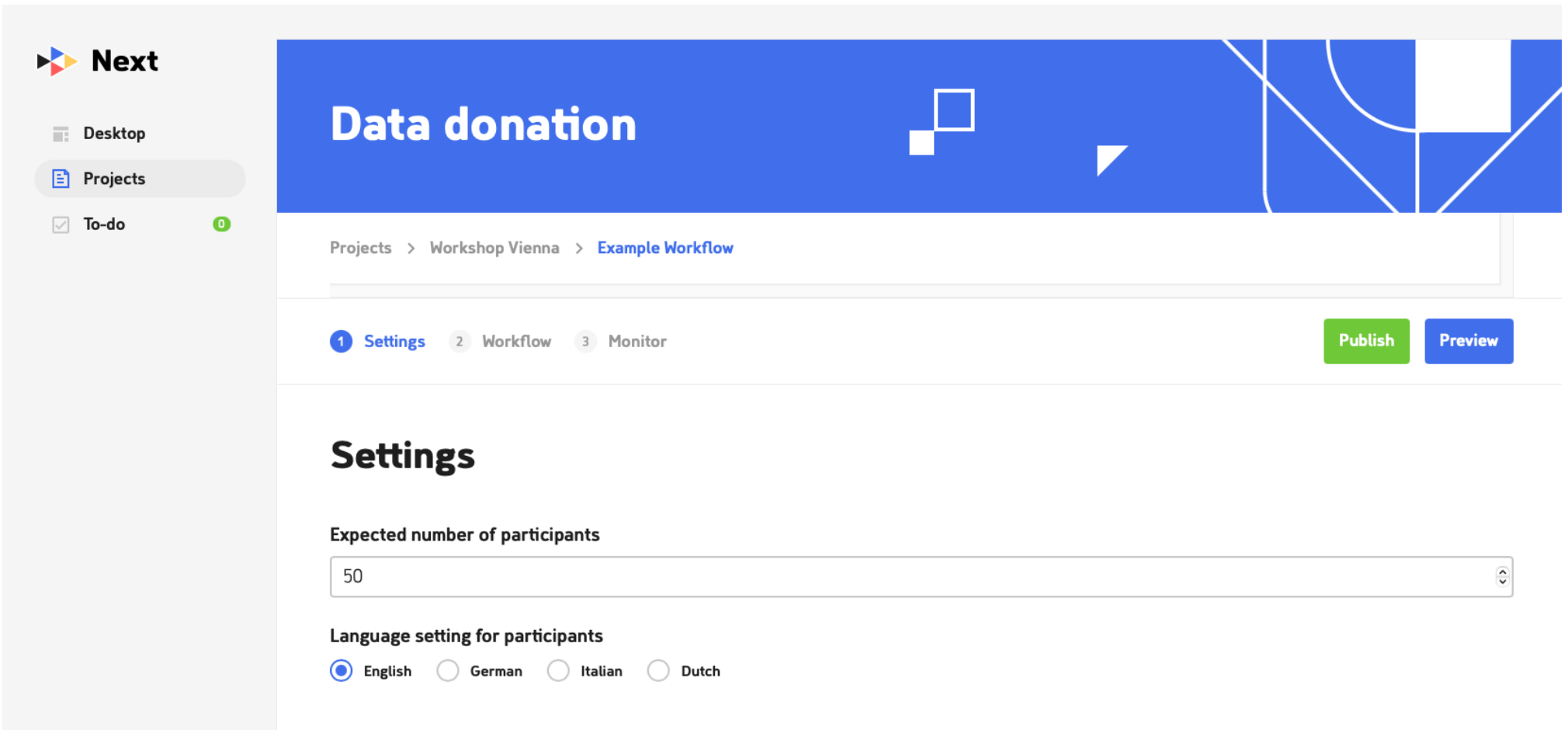
Choose a tool, e.g., ...

- **Next** (Boeschoten et al., 2023) (different measurements, different platforms)
- **Data** Donation Module (Pfiffner et al., 2022) (different measurements, different platforms)
- **Dona** (Hakobyan et al., 2025) (messaging data, different platforms)
- **WhatsR** (Kohne & Montag, 2024) (messaging data, WhatsApp)

Step I.II: Which data donation tool do I use?

Choose a tool, e.g., ...

- **Next** (Boeschoten et al., 2023) (different measurements, different platforms)



Step I.II: Which data donation tool do I use? 

Relevant questions include...

- Can I/do I want to change extraction scripts myself? (example: TikTok)
- Can I/do I want to provide my scripts to other researchers?
- Can I/do I have to host data on my own server?

Step I.II: Which data donation tool do I use? 🖥️

Key questions:

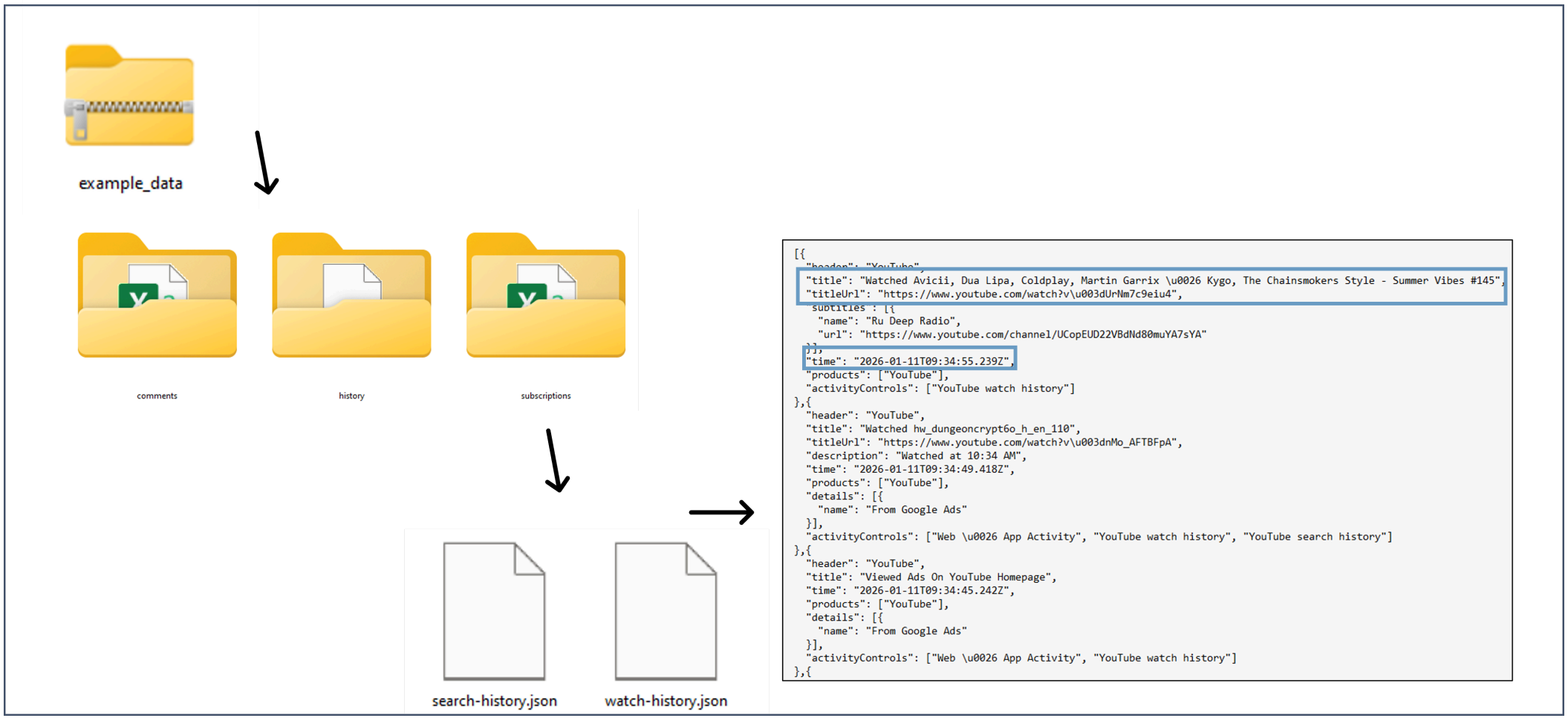
- **Which data donation tool do I use?** 🖥️
- Which variables do I extract? 🔎
- How do I anonymize data? 🙈

Step I.II: Which variables do I extract? 🔎

Key questions:

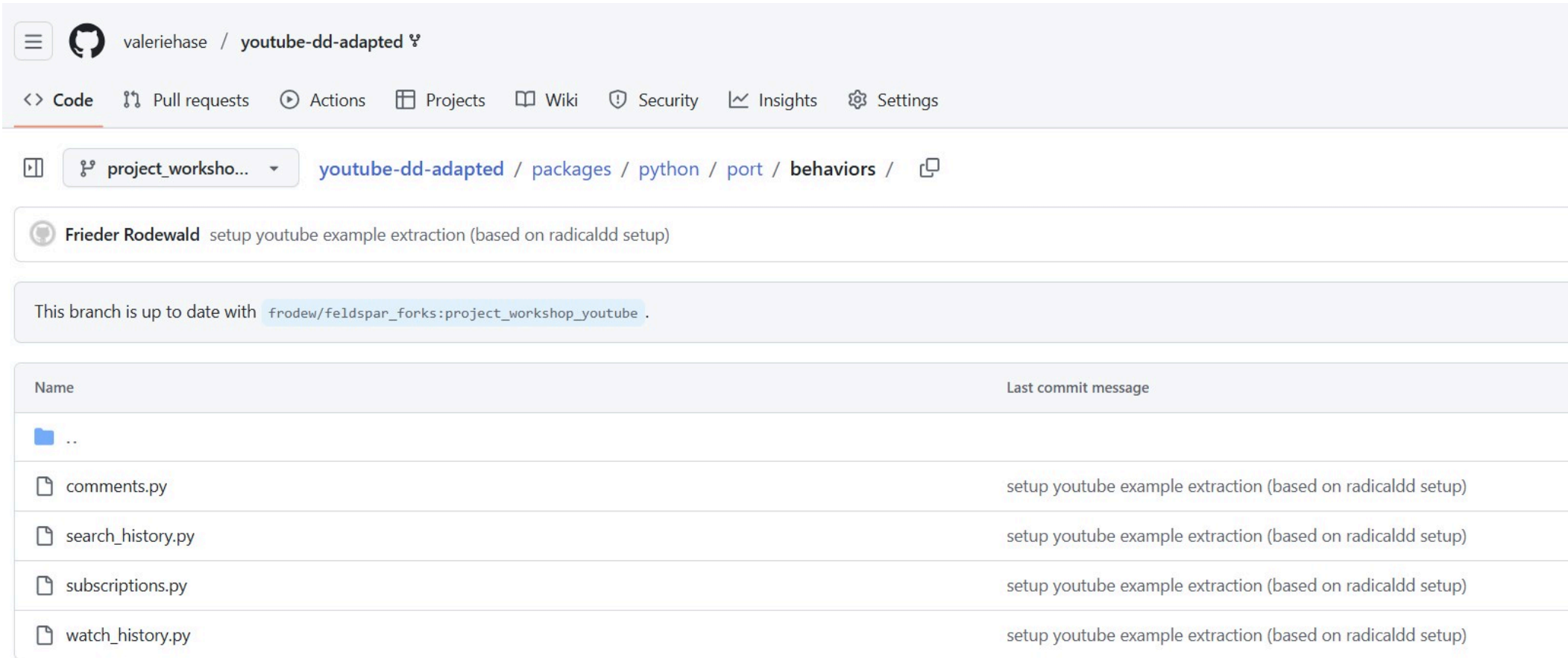
- Which data donation tool do I use? 🖥️
- **Which variables do I extract?** 🔎
- How do I anonymize data? 🙈

Step I.II: Which variables do I extract?



Step I.II: Which variables do I extract?

You can find the following Python code for data extraction here:



Step I.II: Which variables do I extract?

Code | Blame | 120 lines (98 loc) · 4.18 KB

```
1   from datetime import datetime
2
3   import pandas as pd
4
5   from port.extraction_helpers import extract_single_file_from_zip
6
7   # Patterns to find the relevant files for this behavior (English and German)
8   patterns = ["history/watch-history", "Verlauf/Wiedergabeverlauf"]
9
10  # Title used in prompt_consent() to describe this behavior
11  title = {
12      "en": "Watch History",
13  }
14
15
16  def format_timestamp(timestamp_str): ...
27          return timestamp_str
28
29
30  def extract_watch_history(zip_file_path): ...
120     return df
```

Step I.II: Which variables do I extract? 

Key decisions include:

- Which files “count” towards measuring my latent concepts?
- Which meta data do I want to extract? (e.g., time stamps, video IDs)
- Do I include multilingual data?

Step I.II: How do I anonymize data? 🙈

Key questions:

- Which data donation tool do I use? 🖥️
- Which variables do I extract? 🔎
- **How do I anonymize data?** 🙈

Please look at your data and discuss: What needs to be anonymized? How could we do this? 🤔

Step I.II: How do I anonymize data? 🙈

Good anonymization may require...

- Whitelists/dictionaries of social media accounts
  - Database of public speakers, Hans-Bredow-Institute (Link)
  - Database of news media and their social media handles, University of Vienna (Link)
- Local, in-tool classification (e.g., topic modeling)
- Manual annotation (e.g., type of contact)
- Aggregation

Step I.II: How do I anonymize data? 🙈

Code | Blame | 1012 lines (1002 loc) · 40.1 KB

```
1   import typing
2   import re
3   from .genuine import unravel_hierarchical_fields
4
5   fb_list_usernames = ['1LIVE',
6                        '12-App',
7                        '20 Minuten',
8                        '3sat',
9                        'Aachener Nachrichten',
10                       'Aachener Zeitung',
11                       'Aarauer Nachrichten',
12                       'Aargauer Zeitung',
13                       'Abendzeitung München',
14                       'Achgut.com - Die Achse des Guten',
15                       'Achtzig - Die Kulturzeitung',
16                       'actu.fr',
17                       'Adpunktum',
18                       'Advantage Wirtschaftsmagazin',
19                       'Aichacher Zeitung',
20                       'Aktuell Obwalden',
21                       'Alfelder Zeitung',
22                       'all-in.de - das Allgäu online.',
23                       'Allgäuer Zeitung',
24                       'Allgemeine Zeitung',
25                       'Allgemeine Zeitung | Coesfeld | Billerbeck | Gescher | Rosendahl | azonline',
```

Figure. Exampe whitelist

Step I.II: How do I anonymize data? 🙈

engagement_timestamp	day	engagement_type	donation_platform	donation_type
2021-12-04 05:41:51	2021-12-04	non-news	Instagram	followed
2021-11-30 13:58:03	2021-11-30	non-news	Instagram	followed
2021-11-26 15:11:16	2021-11-26	non-news	Instagram	followed
2021-11-22 22:00:22	2021-11-22	news	Instagram	followed
2021-11-19 15:22:43	2021-11-19	non-news	Instagram	followed
2021-11-08 16:13:18	2021-11-08	news	Instagram	followed
2021-11-07 15:56:43	2021-11-07	non-news	Instagram	followed
2021-11-01 07:25:09	2021-11-01	non-news	Instagram	followed

Figure. Example anonymized data

⚠ Anonymized does not mean anonymous!

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

Let's have a look at the technical set-up 💻 :

- the platform: https://github.com/eyra/mono
- the data donation tool: https://github.com/eyra/feldspar

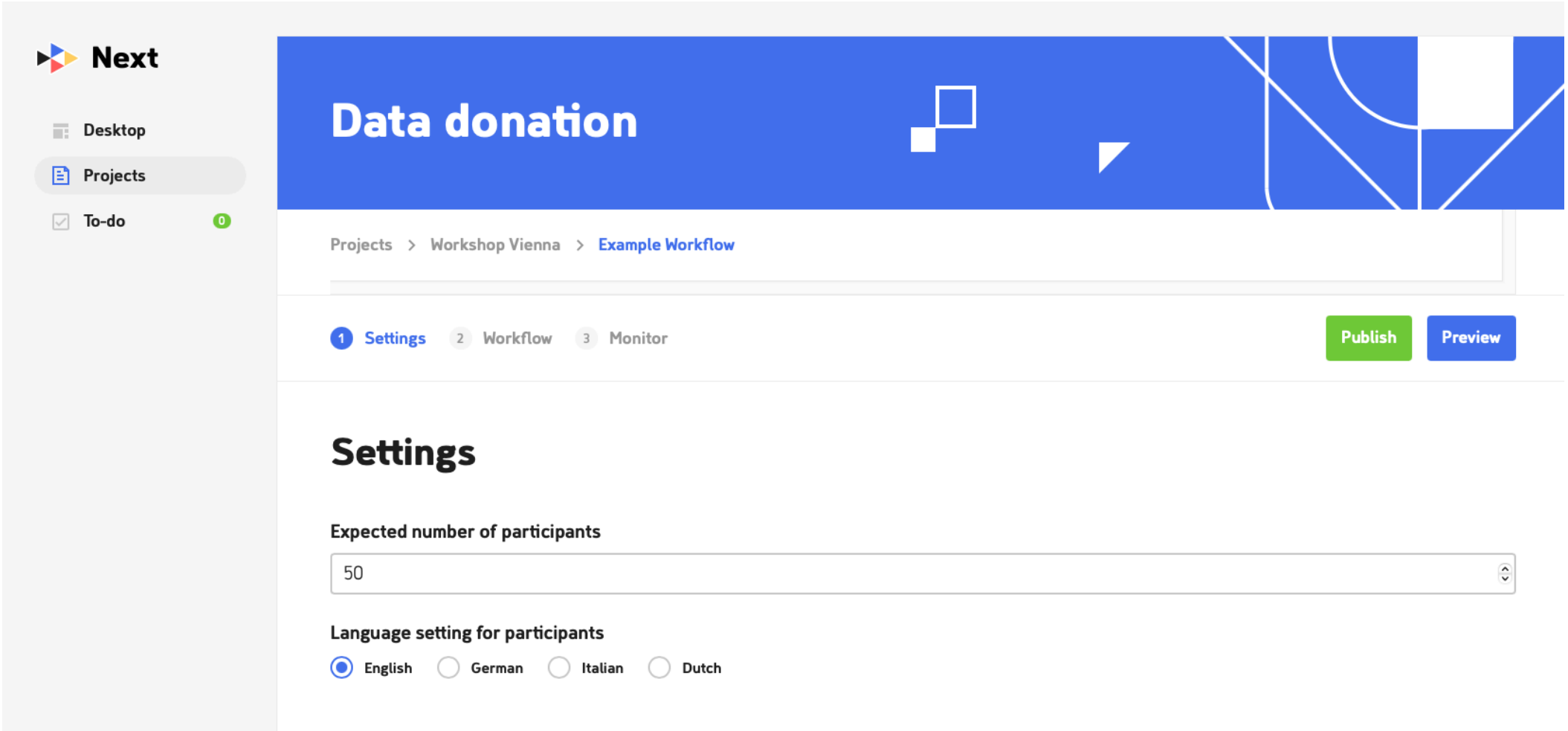


Figure. Next setup

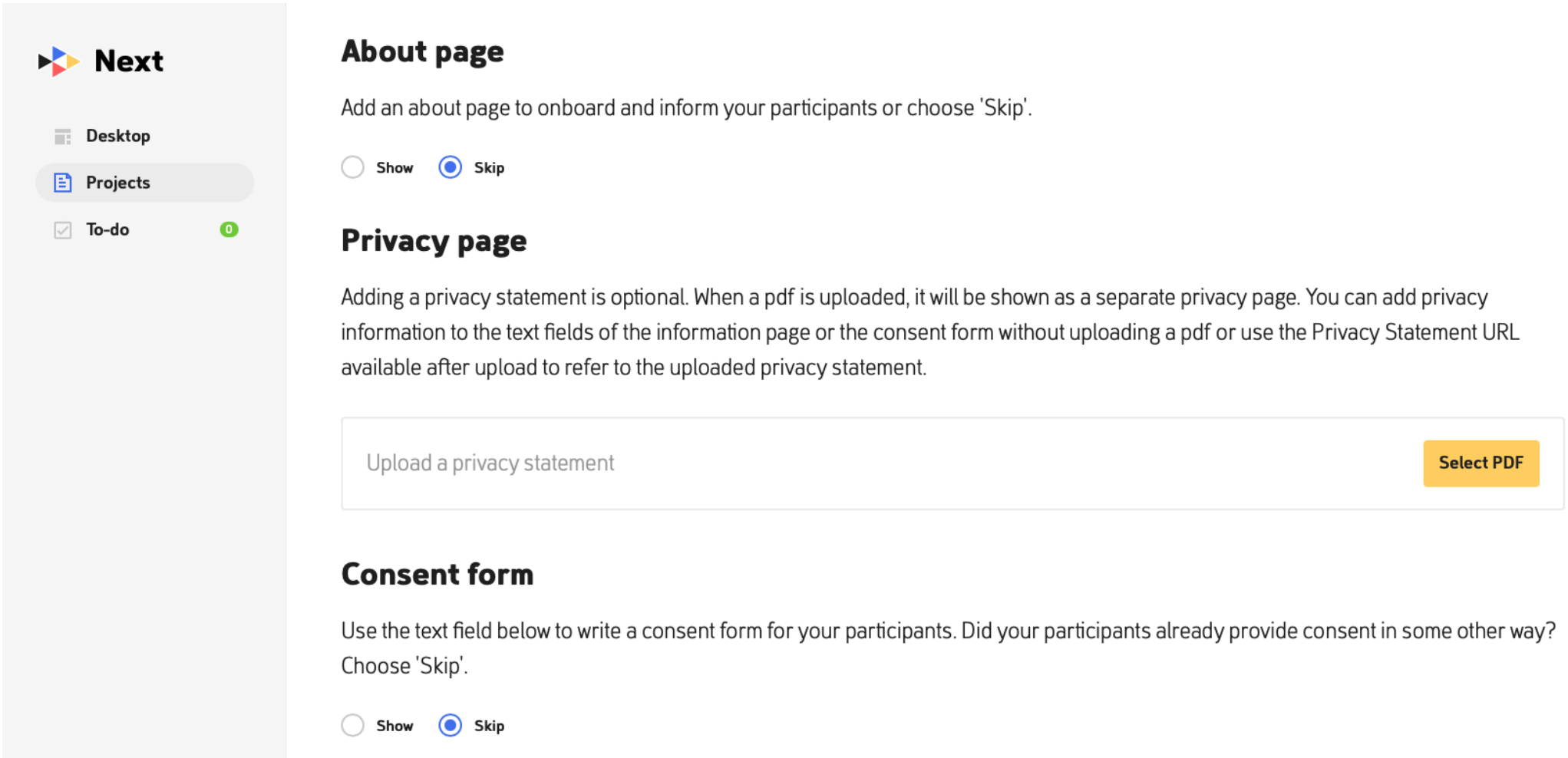


Figure. Next setup

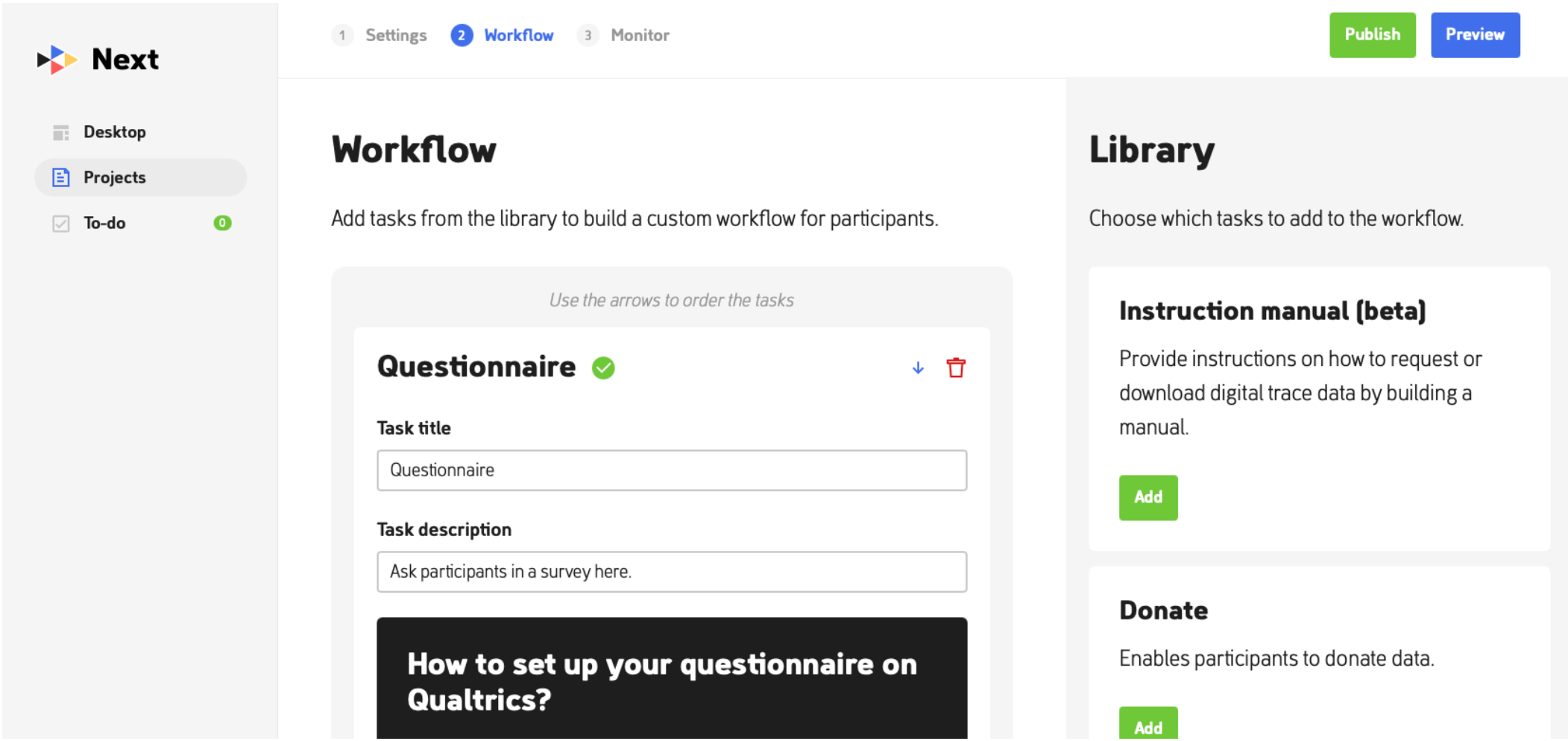


Figure. Next setup

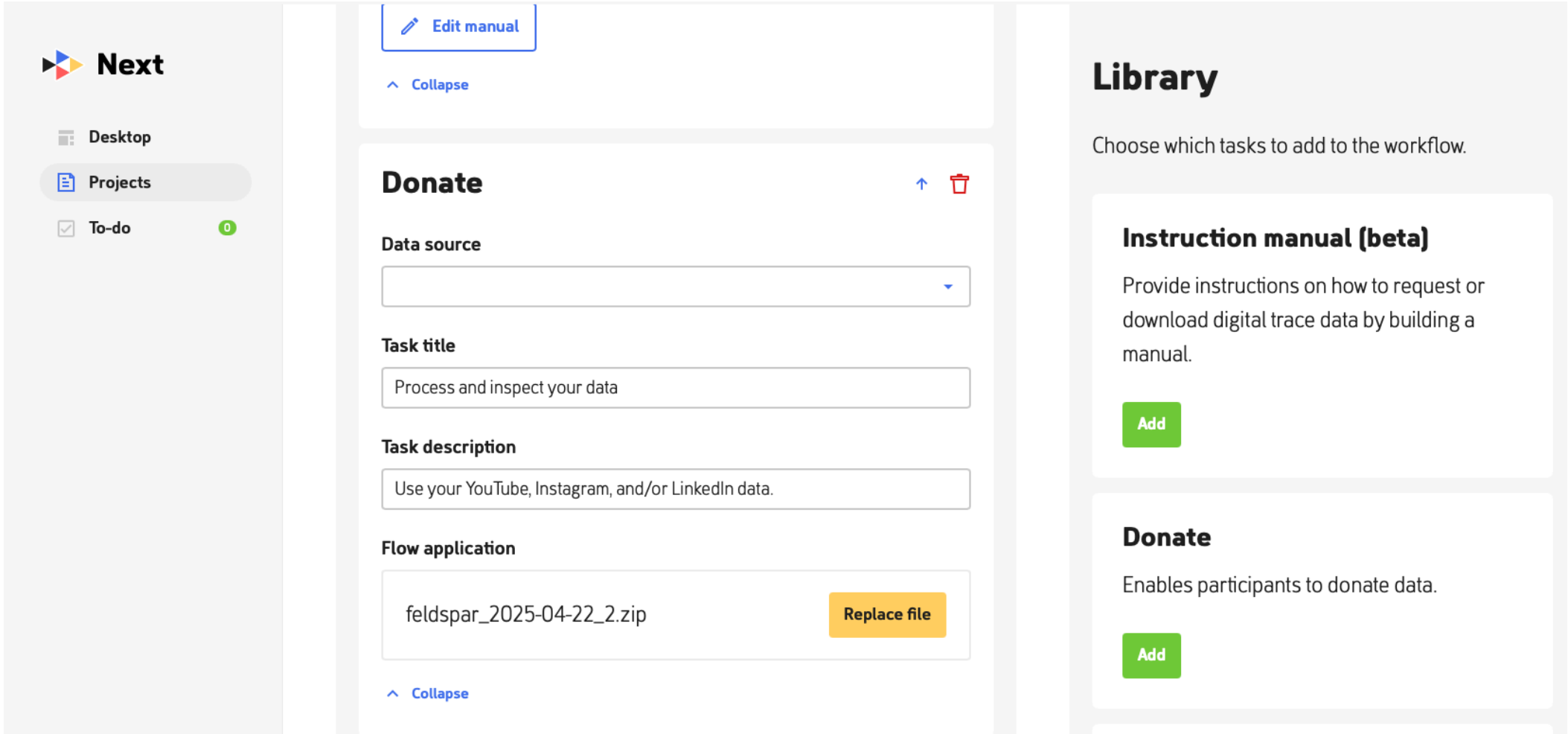


Figure. Next setup

Step I: Research design & tool set-up

1 Research Design & Tool Set-Up

2 Data Cleaning & Augmentation

3 Modelling

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables?

1.3 How do I integrate the tool in surveys & recruit participants?

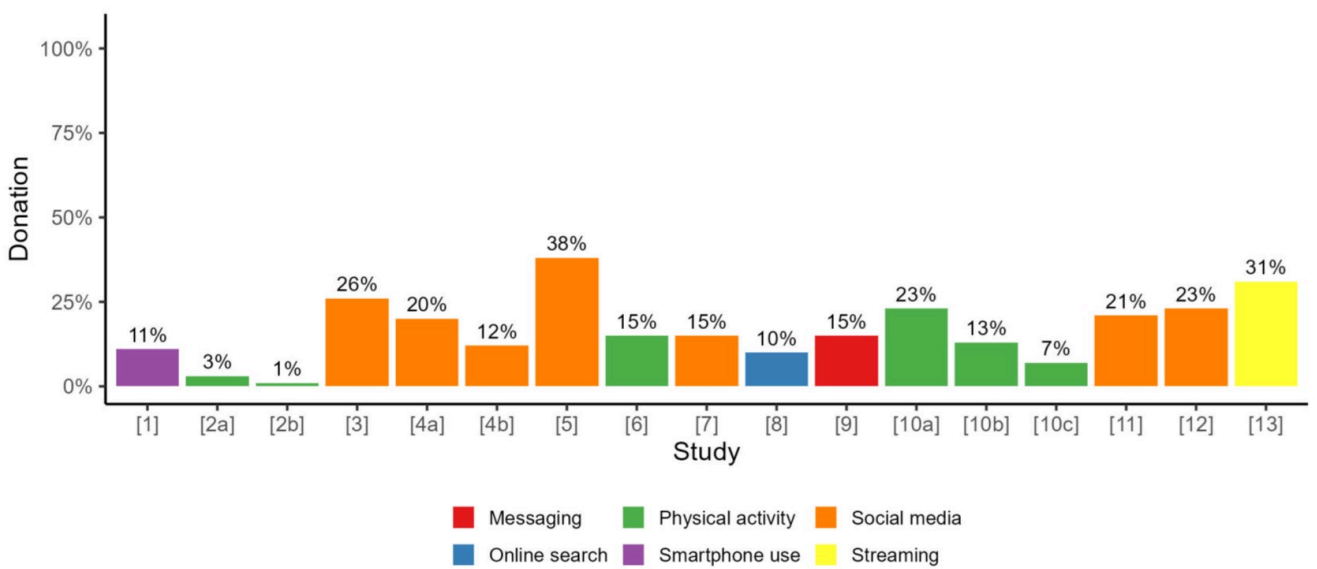
What are (dis-)advantages of online access opt-in panels for data donation? 🤔

Step I.III: How do I integrate the tool in surveys & recruit participants?

- Use a single platform for survey & data donation
- Think about characteristics of your population, e.g.,
  - Who is using the platform where data should be donated from?
  - Who is willing and able to share their data?
  - Can you incentive participants in a meaningful way?
- ⚠️ Often, the main goal will not (or cannot) be to reach a “representative” sample.

Step I.III: How do I integrate the tool in surveys & recruit participants?

- Low response rates (e.g., Hase & 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 (38-1% of survey respondents)
  - Well known intention-behavior gap (Kmetty & Stefkovics, 2025)



Step I: Research design & tool set-up

1 Research Design & Tool Set-Up

2 Data Cleaning & Augmentation

3 Modelling

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables via my data donation tool?

1.3 How do I integrate the tool in surveys & recruit participants?

Figure. Data donation study - researcher perspective

Step II: Data cleaning & augmentation

1 Research Design & Tool Set-Up

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables?

1.3 How do I integrate the tool in surveys & recruit participants?

2 Data Cleaning & Augmentation

2.1 How do I clean and extend data?

2.2 How do I check for bias?

3 Modelling

Figure. Data donation study - researcher perspective

Step II.I: How do I clean and extend data?

This is how your data may look like:

	id	submission_id	filename	n_deleted	insert_timestamp	update_timestamp	entry
7868	308142	5345	liked_posts.json	0	2022-12-09 10:37:45.458707+00:00	2022-12-09 10:37:45.458714+00:00	{"string_list_data":[{"timestamp":1654035032}],"title":"<user>"}
7869	308143	5345	liked_posts.json	0	2022-12-09 10:37:45.458731+00:00	2022-12-09 10:37:45.458737+00:00	{"string_list_data":[{"timestamp":1654034499}],"title":"<user>"}
7870	308144	5345	liked_posts.json	0	2022-12-09 10:37:45.458754+00:00	2022-12-09 10:37:45.458761+00:00	{"string_list_data":[{"timestamp":1654034341}],"title":"<user>"}
7871	308145	5345	liked_posts.json	0	2022-12-09 10:37:45.458777+00:00	2022-12-09 10:37:45.458784+00:00	{"string_list_data":[{"timestamp":1654020807}],"title":"<user>"}
7872	308146	5345	liked_posts.json	0	2022-12-09 10:37:45.458801+00:00	2022-12-09 10:37:45.458808+00:00	{"string_list_data":[{"timestamp":1654020127}],"title":"<user>"}
7873	308147	5345	liked_posts.json	0	2022-12-09 10:37:45.458824+00:00	2022-12-09 10:37:45.458831+00:00	{"string_list_data":[{"timestamp":1654020057}],"title":"tagesschau"}
7874	308148	5345	liked_posts.json	0	2022-12-09 10:37:45.458847+00:00	2022-12-09 10:37:45.458854+00:00	{"string_list_data":[{"timestamp":1654019851}],"title":"<user>"}
7875	308149	5345	liked_posts.json	0	2022-12-09 10:37:45.458871+00:00	2022-12-09 10:37:45.458878+00:00	{"string_list_data":[{"timestamp":1654019739}],"title":"<user>"}
7876	308150	5345	liked_posts.json	0	2022-12-09 10:37:45.458894+00:00	2022-12-09 10:37:45.458901+00:00	{"string_list_data":[{"timestamp":1654019708}],"title":"<user>"}
7877	308151	5345	liked_posts.json	0	2022-12-09 10:37:45.458918+00:00	2022-12-09 10:37:45.458925+00:00	{"string_list_data":[{"timestamp":1653940335}],"title":"<user>"}
7878	308152	5345	liked_posts.json	0	2022-12-09 10:37:45.458941+00:00	2022-12-09 10:37:45.458948+00:00	{"string_list_data":[{"timestamp":1653938012}],"title":"<user>"}
7879	308153	5345	liked_posts.json	0	2022-12-09 10:37:45.458965+00:00	2022-12-09 10:37:45.458971+00:00	{"string_list_data":[{"timestamp":1653937848}],"title":"<user>"}
7880	308154	5345	liked_posts.json	0	2022-12-09 10:37:45.458988+00:00	2022-12-09 10:37:45.458995+00:00	{"string_list_data":[{"timestamp":1653937307}],"title":"<user>"}
7881	308155	5345	liked_posts.json	0	2022-12-09 10:37:45.459011+00:00	2022-12-09 10:37:45.459018+00:00	{"string_list_data":[{"timestamp":1653808843}],"title":"<user>"}
7882	308156	5345	liked_posts.json	0	2022-12-09 10:37:45.459035+00:00	2022-12-09 10:37:45.459042+00:00	{"string_list_data":[{"timestamp":1653781269}],"title":"<user>"}
7883	308157	5345	liked_posts.json	0	2022-12-09 10:37:45.459058+00:00	2022-12-09 10:37:45.459065+00:00	{"string_list_data":[{"timestamp":1653753711}],"title":"sz"}
7884	308158	5345	liked_posts.json	0	2022-12-09 10:37:45.459082+00:00	2022-12-09 10:37:45.459089+00:00	{"string_list_data":[{"timestamp":1653691455}],"title":"<user>"}
7885	308159	5345	liked_posts.json	0	2022-12-09 10:37:45.459105+00:00	2022-12-09 10:37:45.459112+00:00	{"string_list_data":[{"timestamp":1653674965}],"title":"<user>"}
7886	308160	5345	liked_posts.json	0	2022-12-09 10:37:45.459128+00:00	2022-12-09 10:37:45.459135+00:00	{"string_list_data":[{"timestamp":1653674398}],"title":"<user>"}

Figure. Donated data - example

Step II.I: How do I clean and extend data?

This is how your data may look like:

	id	submission_id	filename	n_deleted	insert_timestamp	update_timestamp	entry
1	708905	9073	Suchverlauf.json	0	2022-12-17 12:43:07.127782+00:00	2022-12-17 12:43:07.127790+00:00	{"title":"Gesucht nach: kinocheck","titleUrl":"https://www.youtube.com/results?search_query=kinocheck"}
2	1050798	10102	Suchverlauf.json	0	2022-12-20 11:08:43.968028+00:00	2022-12-20 11:08:43.968035+00:00	{"title":"Gesucht nach: anno 1602 denkmal","titleUrl":"https://www.youtube.com/results?search_query=anno+1602+denkmal"}
3	619493	8665	Suchverlauf.json	0	2022-12-16 21:04:58.414825+00:00	2022-12-16 21:04:58.414832+00:00	{"title":"Gesucht nach: ytitti","titleUrl":"https://www.youtube.com/results?search_query=ytitti"}
4	938862	9908	Suchverlauf.json	0	2022-12-19 13:26:30.762649+00:00	2022-12-19 13:26:30.762657+00:00	{"title":"Coop Erbjudande v6 angesehen","titleUrl":"https://www.youtube.com/watch?v=ql1goWZD8nQ"}
5	1289477	10178	Suchverlauf.json	0	2022-12-28 15:33:30.872355+00:00	2022-12-28 15:33:30.872362+00:00	{"title":"The spring collection angesehen","titleUrl":"https://www.youtube.com/watch?v=fi49A9iB1hA"}

Figure. Donated data - example

Step II.I: How do I clean and extend data?

Often, we need to further preprocess collected data through...

- Manual annotation (by participants or researchers)
- APIs/scraping to extend collected data
- Text-as-data methods

Step II: Data cleaning & augmentation

1 Research Design & Tool Set-Up

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables?

1.3 How do I integrate the tool in surveys & recruit participants?

2 Data Cleaning & Augmentation

2.1 How do I clean and extend data?

2.2 How do I check for bias?

3 Modelling

Figure. Data donation study - researcher perspective

Step II: Data cleaning & augmentation

1 Research Design & Tool Set-Up

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables?

1.3 How do I integrate the tool in surveys & recruit participants?

2 Data Cleaning & Augmentation

2.1 How do I clean and extend data?

2.2 How do I check for bias?

3 Modelling

Figure. Data donation study - researcher perspective

Step II.II: How do I check for bias?

- Errors in representation and measurements
  - based on systematic drop-out (Pak et al., 2022)
  - based on systematic misclassification of digital traces (TeBlunthuis et al., 2024)

👉 You know the drill: We will talk about this in session 4️⃣.

Step III: Modelling

1 Research Design & Tool Set-Up

1.1 Which theoretical questions do I want to answer?

1.2 How do I operationalize key variables?

1.3 How do I integrate the tool in surveys & recruit participants?

2 Data Cleaning & Augmentation

2.1 How do I clean and extend data?

2.2 How do I check for bias?

3 Modelling

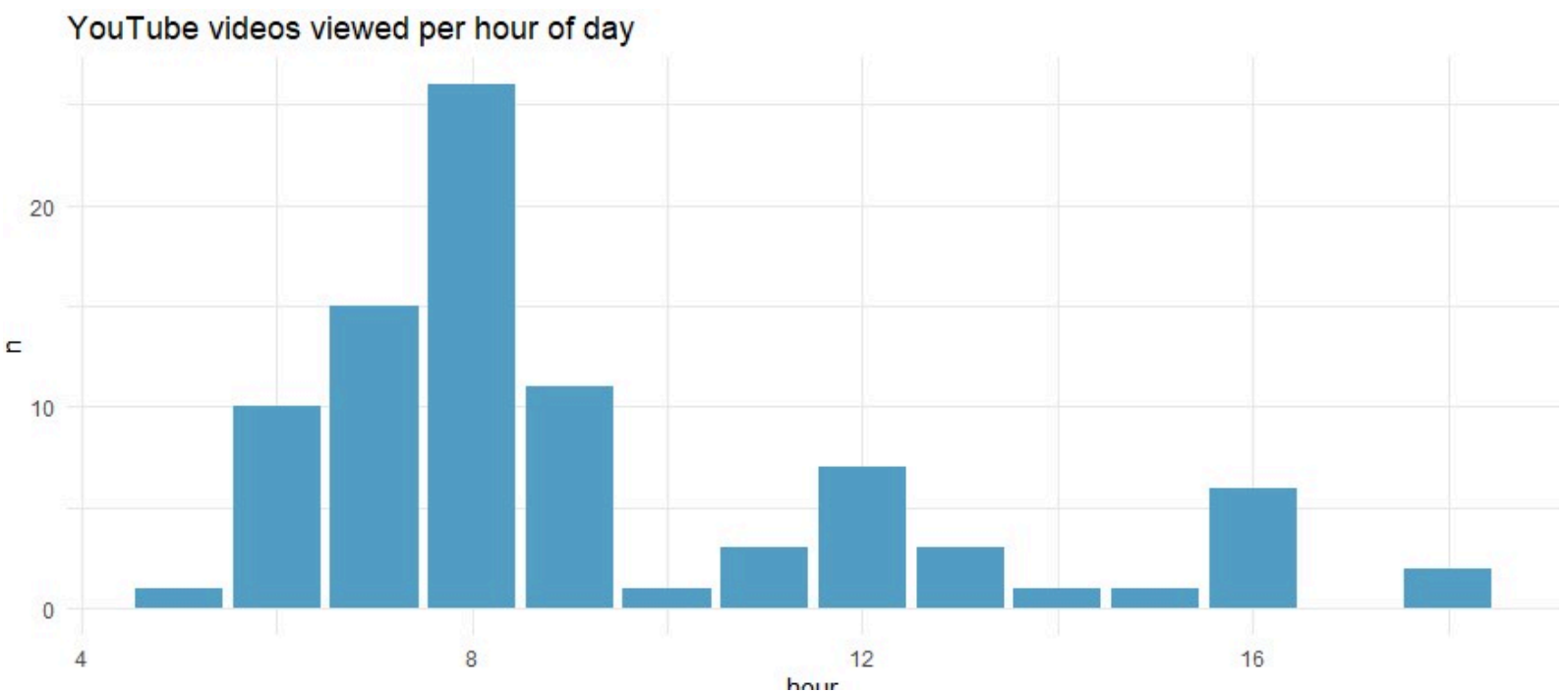
3. How do I analyze results?

Figure. Data donation study - researcher perspective

📢 Task 3: Example Analysis of YouTube Watch history

Download the "Data for Task 3" from the website or use your own YouTube watch history. Also, load the respective R-code. Run the code (you just have to change the location and name of your data).

1. On which day do you mostly watch YouTube?
2. At what time do you mostly watch YouTube?



The idea for this code and analysis was provided by Michael Scharkow, University of Mainz.

Step III.I: How do I analyze results?

For inferential modeling, consider (Clemm Von Hohenberg et al., 2024)....

- Creating indices from different metrics (e.g., liking, sharing, or commenting) and testing their consistency
- Accounting for missing data (units, measures)
- Skewed data (e.g., zero-inflation) and non-linear relationships
- Hierarchical structure (nested in participants, metrics, platforms) and within/between variance
- Advanced longitudinal approaches (e.g., with respect to sequences, feedback loops and causal inference)

Summary: Researcher perspective 

- **Summary**: Key steps include...
    1. Research design & tool set-up
    2. Data cleaning & augmentation
    3. Modelling
- **Further literature**:
    - Boeschoten et al. (2022)
    - Carrière et al. (2024)

Questions? 

References

Boeschoten, L., Mendrik, A., Van Der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., & Oberski, D. L. (2022). Privacy-preserving local analysis of digital trace data: A proof-of-concept. *Patterns*, *3*(3), 100444. https://doi.org/10.1016/j.patter.2022.100444

Boeschoten, L., Schipper, N. C. de, Mendrik, A. M., Veen, E. van der, Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, *8*(90), 5596.

Brinberg, M., & Ram, N. (2021). Do New Romantic Couples Use More Similar Language Over Time? Evidence from Intensive Longitudinal Text Messages. *Journal of Communication*, *71*(3), 454–477. https://doi.org/10.1093/joc/jqab012

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

Clemm Von Hohenberg, B., Stier, S., Cardenal, A. S., Guess, A. M., Menchen-Trevino, E., & Wojcieszak, M. (2024). Analysis of Web Browsing Data: A Guide. *Social Science Computer Review*, *42*(6), 1479–1504. https://doi.org/10.1177/08944393241227868

Corten, R., Boeschoten, L., Carrière, T., Jongerius, S., Struminskaya, B., Mulder, J., Zahedi, P., Nadi Najafabadi, S., & Mendrik, A. (2025). Assessing mobile instant messenger networks with donated data. *Social Network Analysis and Mining*. https://doi.org/10.1007/s13278-025-01550-8

Hakobyan, O., Hillmann, P.-J., Martin, F., Böttinger, E., & Drimalla, H. (2025). Development and evaluation of Dona, a privacy-preserving donation platform for messaging data from WhatsApp, Facebook, and Instagram. *Behavior Research Methods*, *57*(3), 94. https://doi.org/10.3758/s13428-024-02593-z

Hase, V., & Haim, M. (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, F., Pankowska, P. K., Cernat, A., & Bach, R. L. (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), 279–293. https://doi.org/10.1177/1525822X231225907

Kmetty, Z., & Stefkovics, Á. (2025). Validating a willingness to share measure of a vignette experiment using real-world behavioral data. *Scientific Reports*, *15*(1), 9319. https://doi.org/10.1038/s41598-025-92349-2

Kohne, J., & Montag, C. (2024). ChatDashboard: A Framework to collect, link, and process donated WhatsApp Chat Log Data. *Behavior Research Methods*, *56*(4), 3658–3684.

Loecherbach, F., Moeller, J., Trilling, D., & Van Atteveldt, W. (2024). What is news? Mapping the diversity of news experiences in digital trace data. *Journalism*, 14648849241303115. https://doi.org/10.1177/14648849241303115

Pak, C., Cotter, K., & Thorson, K. (2022). Correcting Sample Selection Bias of Historical Digital Trace Data: Inverse Probability Weighting (IPW) and Type II Tobit Model. *Communication Methods and Measures*, *16*(2), 134–155. https://doi.org/10.1080/19312458.2022.2037537

Pfiffner, N., Witlox, P., & Friemel, T. N. (2022). *Data Donation Module*. https://github.com/uzh/ddm

Pierce-Grove, R., & Watkins, E. A. (2024). Integrating trace data into interviews: Better interviews, better data. *Convergence: The International Journal of Research into New Media Technologies*, *30*(6), 2059–2074. https://doi.org/10.1177/13548565241300897

TeBlunthuis, N., Hase, V., & Chan, C.-H. (2024). Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can! *Communication Methods and Measures*, *18*(3), 278–299. https://doi.org/10.1080/19312458.2023.2293713

Virtanen, M. T., Vepsäläinen, H., & Koivisto, A. (2021). Managing several simultaneous lines of talk in Finnish multi-party mobile messaging. *Discourse, Context & Media*, *39*, 100460. https://doi.org/10.1016/j.dcm.2020.100460

Wojcieszak, M., Chang, R.-C. (Anna)., & Menchen-Trevino, E. (2023). Political content and news are polarized but other content is not in YouTube watch histories. *Journal of Quantitative Description: Digital Media*, *3*. https://doi.org/10.51685/jqd.2023.018

Xuanjun Gong, & Richard Huskey. (2023). Media selection is highly predictable, in principle. *Computational Communication Research*, *5*(1), 1. https://doi.org/10.5117/CCR2023.1.15.GONG

Yu, X., Haroon, M., Menchen-Trevino, E., & Wojcieszak, M. (2024). Nudging recommendation algorithms increases news consumption and diversity on YouTube. *PNAS Nexus*, *3*(12), pgae518. https://doi.org/10.1093/pnasnexus/pgae518