

Follow the User?!

Data Donation Studies for Collecting Digital Trace Data

Session **3**: Data Donation Studies (Researcher Perspective)

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



Part of the SPP DFG Project [Integrating Data Donations in Survey Infrastructure](#)

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

Data donation study - researcher perspective



Figure. Data donation study - researcher perspective

Agenda

1. Research design & tool set-up
2. Data cleaning & augmentation, including
 - Task 3: Classify search terms
3. Modelling digital traces



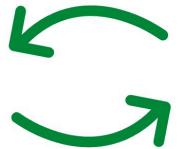
Image by Hope House Press via Unsplash

1) Research design & tool set-up (Frieder)



Source: Image by Markus Winkler via Unsplash

Step I: Research design & tool set-up



1

Research Design & Tool Set-Up

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?

2

Data Cleaning & Augmentation

3

Modelling

Figure. Data donation study - researcher perspective

Step I: Research design & tool set-up

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?

Step I: Research design & tool set-up

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?

Step I.1 Which questions do I want to answer?

This may sound silly but:

- Novel method, few empirical applications
- To date: methodological playground
- *What good is a method that is not used to advance theories/empirical knowledge?*

Step I: Research design & tool set-up

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?

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

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

- Port ([Boeschoten et al., 2023](#)) (Netherlands, different platforms)
- Data Donation Module ([Pfiffner et al., 2022](#)) (Switzerland, different platforms)
- WhatsR ([Kohne & Montag, 2024](#)) (Germany, WhatsApp)

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

- Participants “upload” data
- Local extraction, anonymization, & aggregation
- Users can delete data
- Informed consent, only then: send to researcher server

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

- Participants “upload” data
- Local **extraction**, anonymization, & aggregation
- Users can delete data
- Informed consent, only then: send to researcher server

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

Extraction:

Name	Typ
ads_and_businesses	Dateiordner
ads_and_topics	Dateiordner
apps_and_websites	Dateiordner
autofill_information	Dateiordner
avatars_store	Dateiordner
comments	Dateiordner
contacts	Dateiordner
content	Dateiordner
device_information	Dateiordner
digital_wallets	Dateiordner
events	Dateiordner
followers_and_following	Dateiordner
fundraisers	Dateiordner
guides	Dateiordner
information_about_you	Dateiordner
likes	Dateiordner
login_and_account_creation	Dateiordner
loyalty_accounts	Dateiordner
media	Dateiordner
media_settings	Dateiordner
messages	Dateiordner

Specific folders & metrics are extracted via CSS



Figure. Filtering data - File extraction

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

Extraction:

Name	Last commit message
..	
api	Merge remote-tracking branch 'upstream/master'
__init__.py	Refactor for extensibility
instagram_extraction_functions.py	add functions for search, messages, time spent and sessions frequency...
instagram_extraction_functions_dict.py	add device usage, search queries, positions, profile and reaction to ...
linkedin_extraction_functions.py	add linkedin function to process saved jobs and change device_usage f...
linkedin_extraction_functions_dict.py	add linkedin function to process saved jobs and change device_usage f...
main.py	Refactor Feldspar component integration and update worker URL: remove...
script.py	Merge remote-tracking branch 'upstream/master'
youtube_extraction_functions.py	remove redundant error checking
youtube_extraction_functions_dict.py	add device usage, search queries, positions, profile and reaction to ...

Figure. Filtering data - Python code

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

Extraction:

```
✓ def extract_ads_seen(ads_seen_json, locale):
    """extract ads_information/ads_and_topics/ads_viewed -> list of authors per day"""

    tl_date = translate("date", locale)
    tl_value = translate(
        {"en": "Seen accounts", "de": "Gesehene Konten", "nl": "Geziene accounts"},
        locale,
    )

    timestamps = [
        t["string_map_data"]["Time"]["timestamp"]
        for t in ads_seen_json["impressions_history_ads_seen"]
    ] # get list with timestamps in epoch format (if author exists)
    dates = [epoch_to_date(t) for t in timestamps] # convert epochs to dates
    authors = [
        i["string_map_data"]["Author"]["value"]
        if "Author" in i["string_map_data"]
        else translate(
            {
                "en": "Unknown account",
                "de": "Unbekanntes Konto",
                "nl": "Onbekend account",
            },
            locale,
        )
        for i in ads_seen_json["impressions_history_ads_seen"]
    ] # not for all viewed ads there is an author!

    adds_viewed_df = pd.DataFrame({tl_date: dates, tl_value: authors})

    aggregated_df = adds_viewed_df.groupby(tl_date)[tl_value].agg(list).reset_index()

    return aggregated_df
```

Figure. Filtering data - Python code

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

- Participants “upload” data
- Local extraction, **anonymization**, & aggregation
- Users can delete data
- Informed consent, only then: send to researcher server

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

Anonymization 🤡:

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',
26                     'Alpenparlament.TV',
27                     'Alpenschau.com',
28                     'Andelfinger Zeitung',
```

Figure. Anonymization - Example of Whitelists

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

Anonymization 🤡:

engagement_timestamp	day	engagement_type	donation_platform	donation_type
2021-12-04 10:37:42	2021-12-04	non-news	Instagram	followed
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 of anonymized data

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

- Participants “upload” data
- Local extraction, anonymization, & aggregation
- Users can delete data
- Informed consent, only then: send to researcher server

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

Aggregation

```
✓ def extract_ads_seen(ads_seen_json, locale):
    """extract ads_information/ads_and_topics/ads_viewed -> list of authors per day"""

    tl_date = translate("date", locale)
    tl_value = translate(
        {"en": "Seen accounts", "de": "Gesehene Konten", "nl": "Geziene accounts"},
        locale,
    )

    timestamps = [
        t["string_map_data"]["Time"]["timestamp"]
        for t in ads_seen_json["impressions_history_ads_seen"]
    ] # get list with timestamps in epoch format (if author exists)
    dates = [epoch_to_date(t) for t in timestamps] # convert epochs to dates
    authors = [
        i["string_map_data"]["Author"]["value"]
        if "Author" in i["string_map_data"]
        else translate(
            {
                "en": "Unknown account",
                "de": "Unbekanntes Konto",
                "nl": "Onbekend account",
            },
            locale,
        )
        for i in ads_seen_json["impressions_history_ads_seen"]
    ] # not for all viewed ads there is an author!

    adds_viewed_df = pd.DataFrame({tl_date: dates, tl_value: authors})

    aggregated_df = adds_viewed_df.groupby(tl_date)[tl_value].agg(list).reset_index()

    return aggregated_df
```

Figure. Aggregation - Python code

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

- Participants “upload” data
- Local extraction, anonymization, & aggregation
- Users can **delete data**
- Informed consent, only then: send to researcher server

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

Data deletion by users **X**:

Ihre YouTube Datenspende

Legen Sie fest, ob Sie die untenstehenden Daten spenden möchten. Überprüfen Sie die Daten sorgfältig und passen Sie sie bei Bedarf an. Mit Ihrer Spende tragen Sie zur zuvor beschriebenen Forschung bei. Vielen Dank im Voraus.

0 Welche Kanäle haben Sie abonniert?

1 Seite

DER SPIEGEL

Abonniert Kanal

DER SPIEGEL

Anpassen  Auswahl löschen

Keine Änderungen

Figure. Data deletion

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

This is how much “fun” testing DDTs is:

The screenshot shows a GitHub Issues page with several open pull requests. A central box highlights a single issue with the text "A single (!) issue" and a sad emoji.

Issues:

- EYRA - Datenspende**
 - Einleitender Text einfacher: "Wir anonymisieren nun Ihre Daten. Sie können diese überprüfen und Ihre Einwilligung geben, bevor Sie Daten mit uns teilen. Die Anonymisierung kann einen Moment dauern — vielen Dank für Ihre Geduld."
 - Kleine Anpassung über Datenspende-Übersicht: Überprüfen Sie die Daten sorgfältig und passen Sie sie bei Bedarf an." zu "Mit "Anpassen" können Sie einzelne Datenpunkte bei Bedarf löschen".
 - ganz generell: Soll das immer "0" sein vor den Datentypen?
- Wie viele Verbindungen haben Sie pro Tag hergestellt und we**
 - ganz generell: Bei LinkedIn/YouTube ist das "pro Tag" Teil der Frage, bei Insta in Klammern dahinter (zB Wie oft haben Sie Instagram geöffnet? [Sitzungen pro Tag]). Letzteres finde ich deutlich besser - auf Instagram anpassen?
 - Ich kann theoretisch zweimal hintereinander Datenpakete hochladen. Ist das ein Problem - überschreibt das Daten oder sind mit meiner ID dann einfach zwei drin?
 - Ich würde den Punkt "Übersicht von zusätzlichen XX Informationen" bei allen Datenspenden rausnehmen. Das sind ja nur Dateien, die fehlen, oder? Ist m.E. für Nutzende sehr verwirrend.
- LinkedIn**
 - Kleine Anpassung Text, wenn man LinkedIn 10 Min Paket hochlädt: "Sie haben das unvollständige Datenpaket hochgeladen, welches LinkedIn bereits nach wenigen Minuten gesendet hat. Für unsere Studie bitten wir Sie, uns das Datenpaket zu spenden, dass Sie normalerweise nach circa 24 Stunden erhalten. Bitte laden Sie dieses vollständige Datenpaket noch."
- Verbindungen**
 - "Wie viele Verbindungen haben Sie pro Tag hergestellt und welche Informationen haben diese?": eher "Mit wie vielen Personen haben Sie sich auf LinkedIn pro Tag vernetzt?" Spaltenübersicht "Anzahl der Verbindungen" -> "Anzahl der neuen Kontakte"
 - Bei dem Datentyp frage ich mich, ob wir die anderen Spalten überhaupt brauchen. Sind m.E. schwer verständlich, zT leer (zB E-Mail ist bei mir immer null), wirken privatscheitigmäßig schwierig ("mit vollständigem Namen") und da anonymisiert genau die gleiche "Nummer" meist wie Anzahl Kontakte. Vielleicht reicht ja wirklich die Anzahl der neuen Kontakte (d.h. nur die erste Spalte)
 - Kontakte nicht nach Datum sortiert

A single (!) issue 😢

Figure. Github issues - Testing the tool

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

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)

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



<https://github.com/eyra/mono>

<https://github.com/eyra/feldspar>

The screenshot shows the Next platform interface. On the left, a sidebar has a 'Next' logo and navigation links: Desktop, Projects (selected), and To-do (with a green badge '0'). The main area is titled 'Data donation' with a blue header. Below it, the breadcrumb navigation shows: Projects > Workshop Vienna > Example Workflow. There are three tabs at the top: 1 Settings (selected), 2 Workflow, and 3 Monitor. On the right are 'Publish' and 'Preview' buttons. The 'Settings' section contains fields for 'Expected number of participants' (set to 50) and 'Language setting for participants' (radio buttons for English, German, Italian, Dutch, where English is selected). The background features abstract white geometric shapes on a blue gradient.

Figure. Next setup 1

The screenshot shows the branding configuration page for a participant workflow. On the left, there's a sidebar with a logo and navigation links: Desktop, Projects (which is selected and highlighted in grey), and To-do (with a green notification badge showing '0'). The main area is titled 'Branding' and contains instructions: 'Use the fields below to customize the header of your participant workflow. Curious to see what this will look like for participants? Click 'Preview' at the top right to check it out.' Below this, there are four sections: 'Title' with the value 'Workshop: Donate your Data'; 'Subtitle' with an empty input field; 'Logo' featuring a circular icon with a stylized black mark and a 'Change logo' button; and 'Header image' featuring a dark thumbnail with small white icons and a 'Change image' button.

Branding

Use the fields below to customize the header of your participant workflow. Curious to see what this will look like for participants? Click 'Preview' at the top right to check it out.

Title

Workshop: Donate your Data

Subtitle

Logo

 Change logo

Header image

 Change image

Figure. Next setup 2

The screenshot shows the 'Next' application interface for study setup. On the left, a sidebar lists 'Desktop', 'Projects' (selected), and 'To-do'. The main area has three sections: 'About page', 'Privacy page', and 'Consent form'. Each section includes a text input field, a 'Select PDF' button, and 'Show' or 'Skip' radio buttons.

About page
Add an about page to onboard and inform your participants or choose 'Skip'.
 Show Skip

Privacy page
Adding a privacy statement is optional. When a pdf is uploaded, it will be shown as a separate privacy page. You can add privacy information to the text fields of the information page or the consent form without uploading a pdf or use the Privacy Statement URL available after upload to refer to the uploaded privacy statement.

Upload a privacy statement

Consent form
Use the text field below to write a consent form for your participants. Did your participants already provide consent in some other way? Choose 'Skip'.
 Show Skip

Figure. Next setup 3

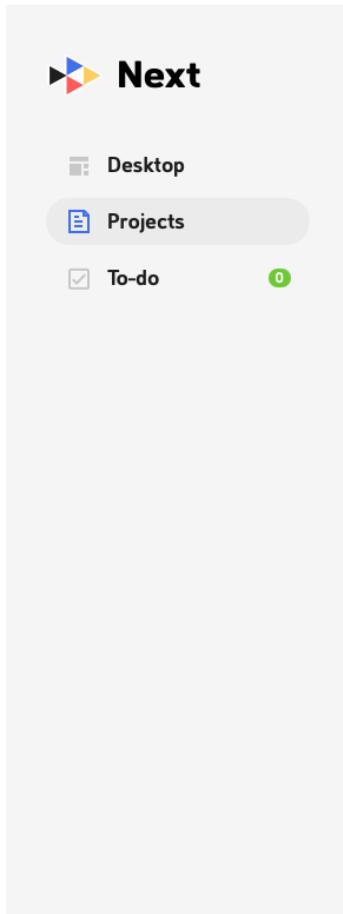


Figure. Next setup 4

Helpdesk page

Use the text field below to inform participants about how to get support when needed. This page will be shown when participants click the 'i' at the right bottom. Check out 'Preview' to see what this will look like. Choose 'Skip' if you are not planning to provide support.

Show Skip

Panel

Connect the panel that you will use for recruiting participants.

Other

 Edit  Disconnect

Copy the url below and provide it to your contact person at the panel agency. With this url participants can visit your data donation workflow. The parameter 'participant' is the unique identifier for a participant.

<https://next.eyra.co/assignment/335/participate?participant={id}> 

Data storage

The data donated by participants is stored on the Next platform.

The screenshot shows the Next platform interface for a 'Data donation' project. The top navigation bar includes a 'Next' logo, a 'Desktop' icon, a selected 'Projects' icon, a 'To-do' icon with a green badge '0', and a search bar placeholder 'Search'. The main header 'Data donation' is displayed on a blue banner above the project details. The project path is 'Projects > Workshop Vienna > Example Workflow'. Below the header, there are three tabs: 'Settings' (grey), 'Workflow' (blue, selected), and 'Monitor' (grey). On the right, there are 'Publish' and 'Preview' buttons. The 'Workflow' section contains a title 'Workflow' and a sub-instruction 'Add tasks from the library to build a custom workflow for participants.' A task card for 'Questionnaire' is shown with a green checkmark, a downward arrow, and a red trash can icon. A note says 'Use the arrows to order the tasks'. The 'Library' section has a title 'Library' and a sub-instruction 'Choose which tasks to add to the workflow.' It also features an 'Instruction manual (beta)' section with a note 'Provide instructions on how to request or download digital trace data by building a manual.'

Figure. Next setup 5

The screenshot shows the Next platform interface. At the top, there is a navigation bar with three tabs: 'Settings' (grey), 'Workflow' (blue, indicating the current active tab), and 'Monitor' (grey). To the right of the tabs are two buttons: 'Publish' (green) and 'Preview' (blue). On the left side, there is a sidebar with icons for 'Desktop', 'Projects' (selected), and 'To-do'. Below the sidebar, the main area has a title 'Workflow' and a subtitle 'Add tasks from the library to build a custom workflow for participants.' A central box contains a task titled 'Questionnaire' with a green checkmark icon. It includes fields for 'Task title' (set to 'Questionnaire') and 'Task description' (set to 'Ask participants in a survey here.'). Below this box is a callout with the text 'How to set up your questionnaire on Qualtrics?'. To the right, there is a 'Library' section with a title 'Library' and a subtitle 'Choose which tasks to add to the workflow.' It lists two items: 'Instruction manual (beta)' and 'Donate', each with an 'Add' button.

1 Settings 2 Workflow 3 Monitor

Publish Preview

Next

Desktop Projects To-do 0

Workflow

Add tasks from the library to build a custom workflow for participants.

Use the arrows to order the tasks

Questionnaire ✓

Task title
Questionnaire

Task description
Ask participants in a survey here.

How to set up your questionnaire on Qualtrics?

Library

Choose which tasks to add to the workflow.

Instruction manual (beta)

Provide instructions on how to request or download digital trace data by building a manual.

Add

Donate

Enables participants to donate data.

Add

Figure. Next setup 6

The image shows a screenshot of the Next platform interface. On the left, there is a sidebar with a 'Next' logo and navigation links: Desktop, Projects (selected), and To-do. The main area displays a task configuration screen for 'Instruction manual (beta)'. The task has a green checkmark icon and edit controls (down, up, delete). It includes fields for 'Data source' (empty dropdown), 'Task title' ('Manual'), and 'Task description' ('Instruct participants on how to request and download their data.'). Below this is a large black callout box with white text containing instructions for setting up participant instructions. The right side shows a 'Library' section with a list of available tasks: 'Instruction manual (beta)' (with a description and 'Add' button), and 'Donate' (with a description and 'Add' button).

Instruction manual (beta) ✓

Data source

Task title

Manual

Task description

Instruct participants on how to request and download their data.

How to set up participant instructions?

- Create **sections**, so participants can navigate to the relevant section.
- Use section **labels** to specify use cases, like *iPhone* or *Android*.
- Provide multiple brief **instructions** within each section, including relevant **screenshots** for clarity.

Library

Choose which tasks to add to the workflow.

Instruction manual (beta)

Provide instructions on how to request or download digital trace data by building a manual.

Add

Donate

Enables participants to donate data.

Add

Figure. Next setup 7

The screenshot shows the Next platform interface. On the left, there's a sidebar with a 'Next' logo and three main menu items: 'Desktop', 'Projects' (which is selected), and 'To-do'. The 'To-do' item has a green circular badge with the number '0'. In the center, there's a 'Edit manual' button and a 'Collapse' button. Below these, a 'Donate' task is being configured. The task has the following fields:

- Data source**: A dropdown menu.
- Task title**: A text input field containing "Process and inspect your data".
- Task description**: A text input field containing "Use your YouTube, Instagram, and/or LinkedIn data."
- Flow application**: A file selection box containing "feldspar_2025-04-22_2.zip" with a "Replace file" button next to it.

At the bottom of the central panel, there's another 'Collapse' button. On the right side of the interface, there's a 'Library' section with the following content:

- Instruction manual (beta)**: A description stating "Provide instructions on how to request or download digital trace data by building a manual." It includes a green "Add" button.
- Donate**: A description stating "Enables participants to donate data." It includes a green "Add" button.

Figure. Next setup 8

Display a menu

The screenshot shows the browser developer tools console with the following log entries:

```

Console opened at 10:46:44 AM

[Running with fake bridge]
[ReactEngine] started
[WorkerProcessingEngine] started
[WorkerProcessingEngine] Received event from worker: - "initialiseDone"
[ReactEngine] received eventType: - "initialiseDone"
[ReactEngine] received: initialiseDone
[FakeBridge] received unknown command: {"__type__": "CommandSystemEvent", "name": "initialized"}
[WorkerProcessingEngine] Received event from worker: - "runCycleDone"
[ReactEngine] received eventType: - "runCycleDone"
[ReactEngine] received: event - Object
[ProcessingWorker] initialise
[ProcessingWorker] loading Pyodide
[ProcessingWorker] loading packages
Loading micropip, packaging, numpy, pandas, python-dateutil, six, pytz
Loaded packaging, micropip, six, pytz, python-dateutil, numpy, pandas
[ProcessingWorker] load port package
[ProcessingWorker] runCycle null
>

```

On the right side of the console, there is a list of function calls with their file paths and line numbers:

- ./src/index.tsx — index.tsx:27
- start — engine.tsx:19
- start — worker_engine.ts:56
- worker_engine.ts:18
- handleEvent — worker_engine.ts:36
- handleEvent — worker_engine.ts:39
- send — fake_bridge.ts:11
- worker_engine.ts:18
- handleEvent — worker_engine.ts:36
- handleEvent — worker_engine.ts:44
- initialise — py_worker.e0eb8245b2b8bf52d3ba.js:82
- startPyodide — py_worker.e0eb8245b2b8bf52d3ba.js:93
- loadPackages — py_worker.e0eb8245b2b8bf52d3ba.js:100
- pyodide.asm.js:9:111848
- pyodide.asm.js:9:112112
- installPortPackage — py_worker.e0eb8245b2b8bf52d3ba.js:105
- runCycle — py_worker.e0eb8245b2b8bf52d3ba.js:37

Auto — Page ⌘

Figure. Next setup 9

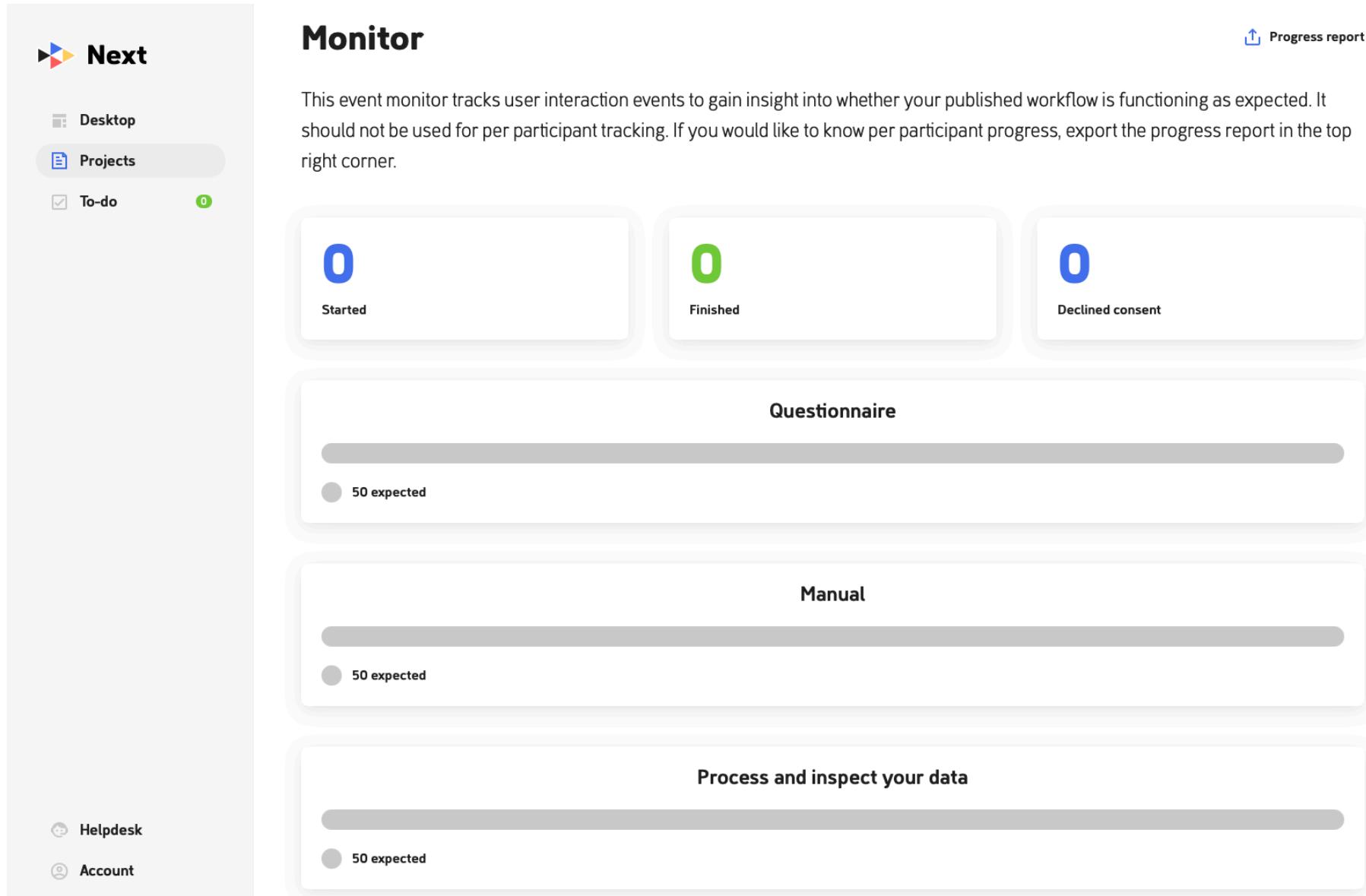


Figure. Next setup 10

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.

Example: Extract list of subscriptions

A	B	C	D
Channel Id	Channel Url	Channel Title	
1 UC0vBXGSyV14uvJ4hECD0l0Q	http://www.youtube.com/channel/UC0vBXGSyV14uvJ4hECD0l0Q	Techquickie	
2 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/ RenÃ© Rebe	
8 UCAXCI-ASTfZqfv9-YklfPIA	http://www.youtube.com/channel/UCAXCI-ASTfZqfv9-YklfPIA	PacKMeN	
9 UCApPPpJ4d3ueW38lArwiWoA	http://www.youtube.com/channel/UCApPPpJ4d3ueW38lArwiWoA	Kenny Beats	
10 UCBa659QWEk1AI4Tg--mrJ2A	http://www.youtube.com/channel/UCBa659QWEk1AI4Tg--mrJ2A	Tom Scott	
11 UCDhu1kICDnf2glev0YbYkDA	http://www.youtube.com/channel/UCDhu1kICDnf2glev0YbYkDA	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_WNNmlpewOtCHf3B0g	http://www.youtube.com/channel/UCJXa3_WNNmlpewOtCHf3B0g	LaurieWired	
16 UCJkMIOu7faDgqh4PfzbPldg	http://www.youtube.com/channel/UCJkMIOu7faDgqh4PfzbPldg	Nerdwriter1	
17 UCMELEMuQqmxTqM4_ArhHPjQ	http://www.youtube.com/channel/UCMELEMuQqmxTqM4_ArhHPjQ	High5	
18 UCMI9UhY1ehLGfOP5KNIKlaQ	http://www.youtube.com/channel/UCMI9UhY1ehLGfOP5KNIKlaQ	Doktor Allwissend	
19 UCMu5gPmKp5av0QCAajKTMhw	http://www.youtube.com/channel/UCMu5gPmKp5av0QCAajKTMhw	ERB	
20 UCN29LJGZ8FY30ysxdTnDsaw	http://www.youtube.com/channel/UCN29LJGZ8FY30ysxdTnDsaw	Filmanalyse	
21 UCNTwGcSEDH1bGhk7l5xFGwA	http://www.youtube.com/channel/UCNTwGcSEDH1bGhk7l5xFGwA	tinseltown	
22 UCOpcACMWbIDls9Z6GERVi1A	http://www.youtube.com/channel/UCOpcACMWbIDls9Z6GERVi1A	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	

Figure. subscriptions.csv

```
1 ...
2     "subscriptions": {
3         "extraction_function": ef.extract_subscriptions,
4         "possible_filenames": ["Abos.csv", "subscriptions.csv"],
5         "title": {
6             "en": "Which channels are you subscribed to?",
7             "de": "Welche Kanäle haben Sie abonniert?",
8             "nl": "Op welke kanalen ben je geabonneerd?",
9         },
10    },
11 ...
```

```
1 def extract_youtube_content_from_zip_folder(zip_file_path, possible_filenames):
2     """Extract content from YouTube data export zip file using filenames"""
3
4     try:
5         with zipfile.ZipFile(zip_file_path, "r") as zip_ref:
6             # Get the list of file names in the zip file
7             filenames = zip_ref.namelist()
8             # Look for matching files
9             for possible_filename in possible_filenames:
10                 for filename in filenames:
11                     if possible_filename in filename:
12                         try:
13                             # Process based on file extension
14                             if filename.endswith(".json"):
15                                 with zip_ref.open(filename) as json_file:
16                                     json_content = json.loads(json_file.read())
17                                     return json_content
18                             elif file_name.endswith(".csv"):
19                                 with zip_ref.open(file_name) as csv_file:
20                                     csv_content = pd.read_csv(csv_file)
21                                     return csv_content
22
23                             # Try the next matching file if there's an error
24                         except Exception as e:
25                             print(f"Error reading file {file_name}: {e}")
26                             continue
27                         # If we've checked all files and found no match
28                         print(f"No file matching file '{possible_filenames}' found")
29                         return None
30             except Exception as e:
31                 print(f"Error extracting YouTube content: {e}")
```

```
1 def extract_subscriptions(subscriptions_csv): # csv file is read before
2     """Extract YouTube channel subscriptions"""
3
4     # Define column name
5     if "Kanaltitel" in subscriptions_csv.columns: # language sensitive
6         channel_column = "Kanaltitel"
7     else:
8         channel_column = "Channel Title"
9
10    # Define description
11    channel_name = "Subscribed Channel"
12
13    # Create DataFrame with just the channel names
14    subscriptions_df = pd.DataFrame({channel_name: subscriptions_csv[channel_column]})
```

0 Which channels are you subscribed to?

< 1 2 3 4 5 6 >

6 pages

Search

Subscribed Channel

Techquickie

Jordan Harrod

Ben Vallack

Wisecrack

Wahre Verbrechen.Wahre Stories

Code Therapy w/ René Rebe

PacKMeN

Adjust

No adjustments

Figure. Processed subscriptions.csv

Step I: Research design & tool set-up

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?**

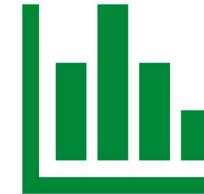
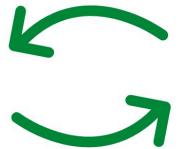
Step I.III: How do I integrate the tool in surveys & recruit participants?

- Often: survey, then forwarding to an external site
- Less often: Integration in existing survey infrastructure ([Haim et al., 2023](#))

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 (37-12% of survey respondents)
 - Well known intention-behavior gap (Kmetty & Stefkovics, 2025)
- Non-response bias
- Primary used in non-probability panels (e.g. online access panels)
- Survey design strategies: For now, 😷 is the only thing that works.
- ➡ Again, we will talk about this in session 4.

Step I: Research design & tool set-up



1

Research Design & Tool Set-Up

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?

2

Data Cleaning & Augmentation

3

Modelling

Figure. Data donation study - researcher perspective

Step II: Data cleaning & augmentation (Valerie)



1 Research Design & Tool Set-Up

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?

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=q1goWZD8nQ"}
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=f49A9IB1hA"}

Figure. Donated data - example

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

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

Task 3: Classify search terms

Download the data for Task 4 from the workshop website. This contains YouTube searches collected from a German social media sample. Either discuss this (no-code group) or do this in R/Python (code group).....

1. How you would clean the data?
2. How you would identify health-related searches using NLP methods?

external_submission_id	search_query	donation_platform
3862	https://www.youtube.com/results?search_query=theorien+d...	YouTube
3862	https://www.youtube.com/results?search_query=Gero+hesse	YouTube
3862	https://www.youtube.com/results?search_query=macarons	YouTube
3862	https://www.youtube.com/results?search_query=Weihnacht...	YouTube
3862	https://www.youtube.com/results?search_query=sallys+welt...	YouTube
9296	https://www.youtube.com/results?search_query=reitmaier	YouTube
9296	https://www.youtube.com/results?search_query=zotero+ma...	YouTube
9296	https://www.youtube.com/results?search_query=einfach+inka	YouTube
9296	https://www.youtube.com/results?search_query=tissot+197...	YouTube
9296	https://www.youtube.com/results?search_query=Druck	YouTube
9272	https://www.youtube.com/results?search_query=der+pate+...	YouTube

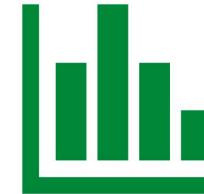
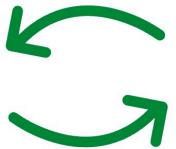
Figure. Donated data - example

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

- Errors in representation and measurements, e.g.
 - 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 4.

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 via my data donation tool?

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 III: Modelling (Valerie)



1 Research Design & Tool Set-Up

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?

2 Data Cleaning & Augmentation

2.1 How do I clean and extend data?

2.2 How do I check for bias?

3 Modelling

3.1 How do I analyze results?

Figure. Data donation study - researcher perspective

Step III.l: How do I analyze results?

Think carefully about...

- How to create indices from different metrics (e.g., liking, sharing, or commenting on content)
- Hierarchical structure (nested in time, metrics, platforms)
- Skewed data, non-linearity

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.
- 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>
- Haim, M., Leiner, D., & Hase, V. (2023). Integrating Data Donations into Online Surveys. *Medien & Kommunikationswissenschaft*, 71(1-2), 130-137. <https://doi.org/10.5771/1615-634X-2023-1-2-130>
- Hase, V., Ausloos, J., Boeschoten, L., Pfiffner, N., Janssen, H., Araujo, T., Carrière, T., De Vreese, C., Haßler, J., Loecherbach, F., Kmetty, Z., Möller, J., Ohme, J., Schmidbauer, E., Struminskaya, B., Trilling, D., Welbers, K., & Haim, M. (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, 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? Data Donation Studies - COMPTEXT - Frieder Rodewald, Valerie Hase