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 1.1 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 🔎:

Specific folders & metrics are extracted via CSS

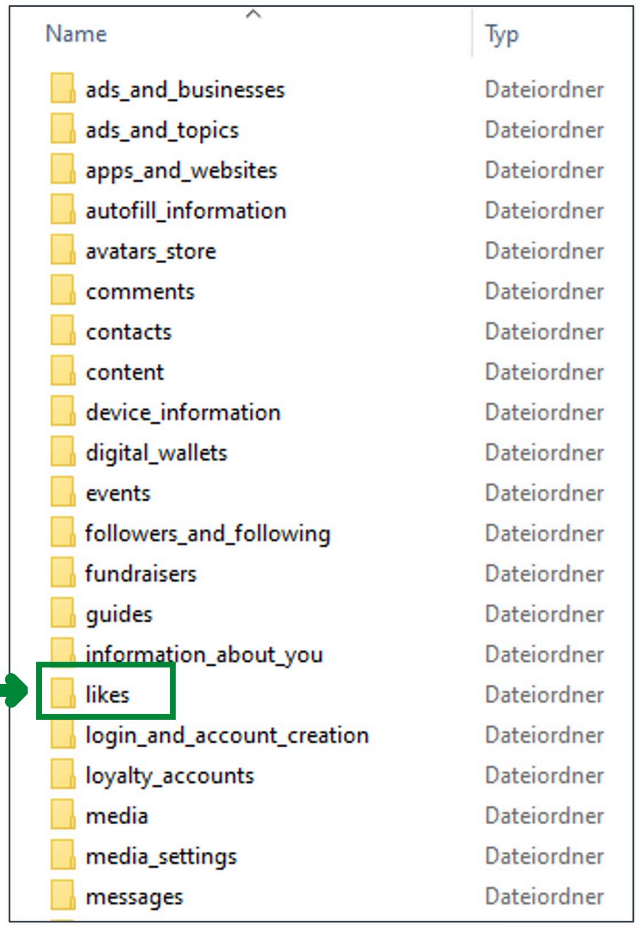

Figure. Filtering data - File extraction

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 🔎:

```python
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

```
import typing
import re
from .genuine import unravel_hierarchical_fields

fb_list_usernames = ['1LIVE',
                     '12-App',
                     '20 Minuten',
                     '3sat',
                     'Aachener Nachrichten',
                     'Aachener Zeitung',
                     'Aarauer Nachrichten',
                     'Aargauer Zeitung',
                     'Abendzeitung München',
                     'Achgut.com - Die Achse des Guten',
                     'Achtzig - Die Kulturzeitung',
                     'actu.fr',
                     'Adpunktum',
                     'Advantage Wirtschaftsmagazin',
                     'Aichacher Zeitung',
                     'Aktuell Obwalden',
                     'Alfelder Zeitung',
                     'all-in.de - das Allgäu online.',
                     'Allgäuer Zeitung',
                     'Allgemeine Zeitung',
                     'Allgemeine Zeitung | Coesfeld | Billerbeck | Gescher | Rosendahl | azonline',
                     'Alpenparlament.TV',
                     'Alpenschau.com',
                     '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 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 🧮:

```python
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 :

Figure. Data deletion

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

This is how much "fun" testing DDTs is:

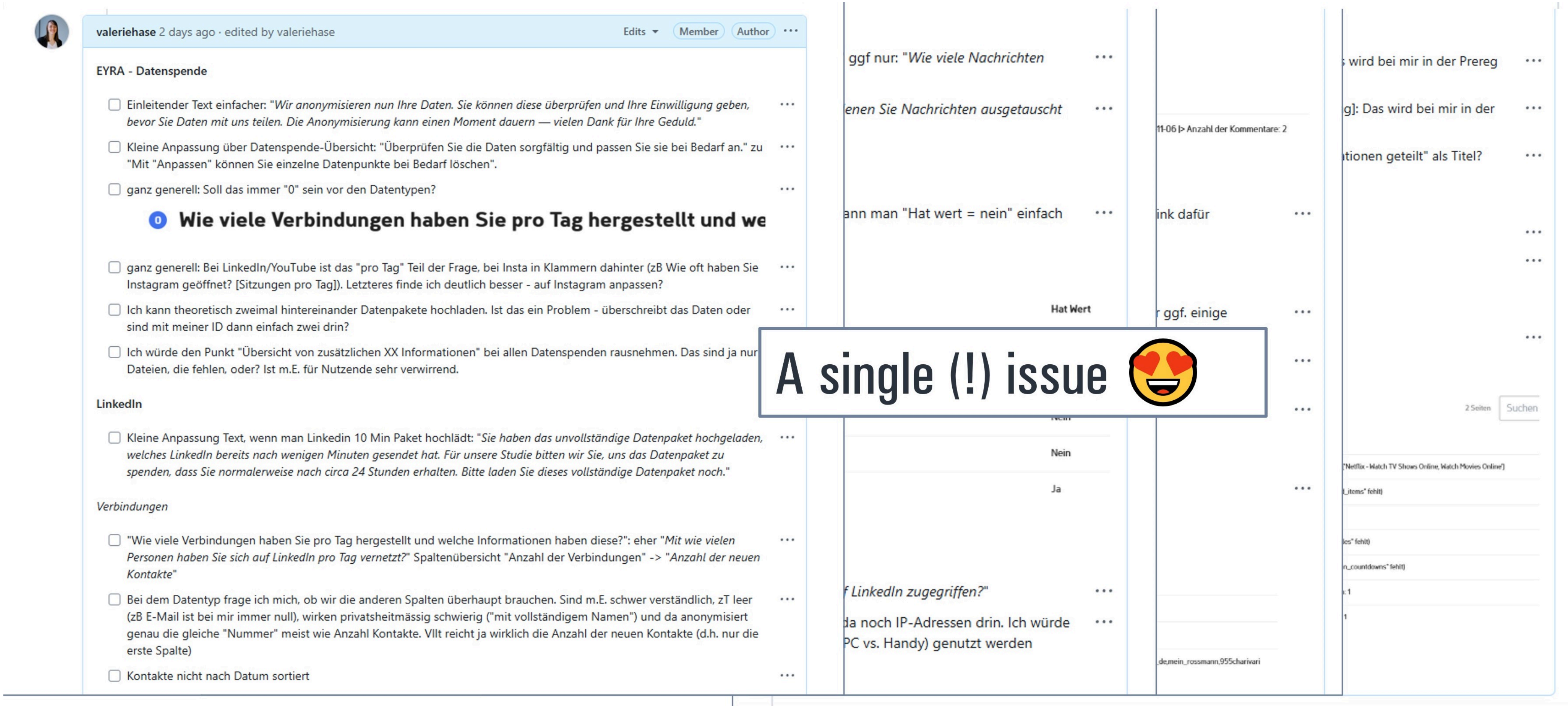

Figure. Github issues - Testing the tool

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

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

Running the DDT locally

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 4️⃣.

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=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=fl49A9iB1hA"}

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.I: 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? 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.

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

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