Anything you would like to share

Evaluating a data donation application in a survey and field study

Welbers, K.; Loecherbach, F.; Lin, Z.; Trilling, D.

DOI
10.5117/CCR2024.2.5.WELB

Publication date
2024

Document Version
Final published version

Published in
Computational Communication Research

Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 6.2 (2024) 1–28
HTTPS://DOI.ORG/10.5117/CCR2024.2.5.WELB

Anything you would like to share: Evaluating a data donation application in a survey and field study

Kasper Welbers
Department of Communication Science, Vrije Universiteit Amsterdam

Felicia Loecherbach
Amsterdam School of Communication Research, University of Amsterdam

Zilin Lin
Amsterdam School of Communication Research, University of Amsterdam

Damian Trilling
Amsterdam School of Communication Research, University of Amsterdam; Department of Journalism Studies, Vrije Universiteit Amsterdamm

Abstract
Data donation methods have shown great potential as a means to measure a person's media consumption behavior and exposure at an unprecedented level of detail. Yet what hampers this potential is that studies often suffer from high drop-out rates, and the accuracy of the digital trace data cannot be taken for granted. To improve the potency of this method, we need to systematically investigate how different recruitment strategies and design choices affect drop-out and accuracy. We used a novel open-source data donation application, and reflect on both a survey and field study where participants were asked to donate their browsing and YouTube history data from Google. Our results confirm that drop-out is high and non-random in the survey study, but adds the positive note that a field lab settings might help alleviate primary barriers of participation. We reflect on opportunities and challenges for data donation research and tools based on log data from our application, questions to participants, and our experience of building the application and guiding users through it.

Keywords: data donation, digital trace data

Introduction

The digital media landscape poses both a challenge and opportunity for measuring a person's media consumption behavior and exposure. On the one hand, classic self-reported measures through survey questions have

become less accurate (Araujo et al., 2017; Prior, 2013; Verbeij et al., 2021). In the time of paper media subscriptions and linear television, we could still ask people about their channels of choice to get a fairly good estimate of what content they were exposed to. But due to the sheer diversity, fragmentation and interconnection of online channels, many now have a hard time recalling where they have been. On the other hand, a large part of online behavior leaves digital traces, and we could potentially collect these traces. This would not only help address the limitations of self-reported exposure measures, but also enable us to measure consumption behavior and exposure on an unprecedented level of detail and from new angles. Digital trace data is observational data that was created in a real-world scenario, instead of created for the purpose of research, which offers new challenges and opportunities for studying human communication and behavior (van Atteveldt & Peng, 2021).

One way to collect these digital traces is by *tracking* participants (Christner et al., 2022; Dvir-Gvirsman, 2017) through software such as browser plug-ins. Despite promising results, social and technological developments– in particular the increasing mobile consumption and the use of proprietary apps–have created many blind spots that are difficult if not impossible to track. An alternative approach instead asks people to *donate* already existing digital traces (Araujo et al., 2022; Menchen-Trevino, 2016). This approach has become especially potent due to developments in privacy regulations, such as the General Data Protection Regulation (GDPR) that empowers individuals to acquire data of their own traces from companies like Google and Meta (Ausloos & Veale, 2020). By guiding and convincing people to request these data and donate them for academic research, we can obtain extensive digital trace data collected by these companies.

In this paper we contribute to *data donation* methodology by discussing two studies in which we asked participants to request their data from Google and donate it to us using a new web application that we developed. The social sciences have gradually built know-how on the use of effective and appropriate incentives, narratives and study designs to recruit participants and improve validity in experiments, surveys, interviews and panels. Despite similarities, data donation methods for collecting participant-centered behavioral traces offer unique challenges, and there remains much to learn regarding when it is a viable method, and how we might lower the barriers for participation (van Driel et al., 2022).

Our contributions are both empirical and methodological. We use two very different study designs (classical online panel survey vs. in-person

collection during a music festival) to gather insights on the motivations to donate and factors that contribute to drop-outs. With this, we extend recent work by Reiss et al. (2022) who also compared different ways of recruiting participants for data donation studies, but did not include an in-person option. These insights can aid researchers in designing future studies by pointing out potential biases and pitfalls that could occur. Our open source tool allows researchers to collect data donations from different data sources in different contexts. Our design took inspiration from the general framework of the WebHistorian application (Menchen-Trevino, 2016) but is specialized for the type of data download packages that major digital platforms like Google and Meta now (are required to) provide (Ausloos & Veale, 2020). In addition to publishing this tool open-source, we also discuss and reflect on our design to inform future tool development.

Data donation

Data donation can be viewed from different perspectives: Firstly, it is a method to collect data – similar to surveys or tracking. Secondly, it is a way to involve participants more in the data collection process. Lastly, it can be used as a means to enhance people's understanding of their own digital footprint. Our application is connected to and builds on many other designs, frameworks, and applications, such as OSD2F (Araujo et al., 2022), PORT (Boeschoten et al., 2023), webhistorian (Menchen-Trevino, 2016) or MIDATA (Shadbolt, 2013).

As a method of data collection, data donation studies can help to collect user-centered digital traces which offer rich insights into behaviors and content exposure. The advantages and disadvantages of this approach in comparison to other data collection methods have been detailed in several overview pieces (Breuer et al., 2022; Ohme et al., 2023; Stier et al., 2020), showing that one main challenge around this collection method is the high burden put on participants and resulting high attrition rates. While the data collected is detailed, non-reactive and can span months to years of digital traces, getting larger samples of participants to take part in the data gathering process through requesting and uploading data remains a main challenge.

This leads to the second aspect: Data donation as way to involve users more in the data collection process – which can both be seen as necessity but also as a way to bring citizens and science closer together. In this regard data donation connects to the broader idea of citizen science, seeing participants as "volunteering" their data for science. In the broader literature of

citizen science, it has been shown that the way studies are presented, the online interfaces that are used to navigate processes, can impact adoption, participation and data quality (Skarlatidou et al., 2019). Furthermore, it has been shown that gamification and insights into data can motivate users to contribute to citizen science projects, especially those who are less motivated by contributing to the public good (Bowser et al., 2013). A larger survey investigated different reasons why people are more or less willing to donate their data to science, showing that – similar to other forms of citizen science – altruistic motives often dominate (Skatova & Goulding, 2019). Social duty is one of the main motivating factors, followed by understanding the purpose of data donation, self-serving motives play only a minor role. Especially the understanding of the purpose of data donation (how and for what is the data being used) can be furthered through explanatory interfaces but also through direct communication with researchers.

This fits the last aspect – increasing understanding of one's own digital footprint. It connects to more general calls that data rights that have been granted to users (e.g., as part of the GDPR legislation or similar laws in Brazil and California) require additional transformation of the data to be useful for end users. Data received from GDPR requests need to be visualized and explained in order to be understandable and useful for users (Schufrin et al., 2020; Veys et al., 2020). There is a need for transparency enhancing tools which inform users which data is being collected, stored and processed in general but also for particular studies (Janic et al., 2013). As part of calls towards a "transparency by design" approach, the idea of a visual turn in making difficult information understandable and more interesting for a wider range of people is strongly encouraged (Rossi & Lenzini, 2020). More specifically for data donation studies ideas such as sequential informed consent (possibility of accepting different data sharing options) and giving a (visual) overview of the data to be shared have been proposed as important interface elements for data donation studies across different fields (Maus et al., 2020). This may also help to convince participants, as a Swiss survey highlights that – among other factors perceived purpose and relevance positively influence the willingness to donate data, while privacy and sensitivity concerns negatively influence it Pfiffner and Friemel (2023). Arguably, applying design principles as outlined above may help alleviate such concerns.

Data donation application

We developed a novel data donation application for this project. Our primary reason for developing a new application instead of using or expanding an existing one was to experiment with design choices that were not easily integrated in the existing tools. By developing a new application, our aim was not only to make a new open-source tool available, but also to contribute to ongoing research into what type of framework and design works best for what purpose.

Application design

We designed our application with four general goals in mind. Firstly, it should not require any installation for participants, and should work on both desktop and mobile phone via the web browser. This facilitates easy integration with survey based recruitment, because participants can directly follow a link to the application. Secondly, the entire process up to the point of donating the data should be *strictly client side*. Participants can explore and filter their data locally on their own device, and we will only be able to see their data *after* they have given informed consent. This makes the informed consent form less complicated, because we can unambiguously say that we don't receive any data that participants did not approve of, and we have also seen this issue come up as a requirement from university ethics boards.

To achieve these first two goals we used the ReactJS (Facebook, 2013) front-end library to build a Single Page Application (SPA). A SPA can look and feel like a native application, but runs directly in a web browser like a regular website. This allowed us to create the entire data donation flow, from importing and parsing the DDPs to visualizing and donating the data, and make it accessible for participants via a simple invitation link.

The third design goal was that it should be easy to create custom import scripts for Data Download Packages (DDPs) (Araujo et al., 2022). This makes it possible for researchers to use the application for multiple platforms, such as the takeout data from Google, Meta and TikTok. But perhaps more importantly, having adjustable DDP import scripts enables researchers to quickly update the data donation flow if the format of a DDP suddenly changes. Even simple changes, such as renaming "URLs" to "Links", could break an import script, and researchers then need to remedy this fast.

The challenge with customizing import scripts in an SPA is that the script needs to be able to run in a web browser. The native language of

the browser is Javascript, but this is not commonly used by computational social scientists. There is an upcoming technology called WebAssembly that makes it possible to compile other languages for the browser (e.g., R, Python), but especially at the start of this project this was still bleeding edge. We therefore decided to explore a third solution, which is to create a simple, standardized set of instructions for parsing the most common data formats used in DDPs, which we called *recipes*.

A recipe itself is a JSON string, and we created a graphical user interface where researchers could upload a DDP and then interactively create or repair the recipe. The benefit of this approach is that it's easy to use, and generates re-usable recipes that can quickly be updated in case a DDP format changes. The primary limitation is that the system needs to be sufficiently flexible to deal with any unforeseen DDP's or changes in the format that can come up. However, we found that a fairly simple system was enough to parse the files that we encountered in DDP's from Google, Meta and TikTok. DDPs mostly stick to common data formats such as CSV, HTML and JSON, that have well established methods for parsing data. CSV was designed to be parsed to a table, HTML has CCS selectors and XPath, and JSON has JSONPath (based on XPath). Our recipe system essentially provides an interface to using these parsers, and works similarly to how spreadsheet software such as Excel has data import wizards (e.g., for parsing CSV files). As our concept and implementation might prove useful for future data donation tools, we published this part of the application separately as an NPM module[1].

Our fourth goal was that it should be possible to include questions about the user's own digital trace data. Firstly, because this allowed us to let participants evaluate the face-validity of their own data, as we discuss more below. Secondly, because having access to someone's digital trace data enables asking personalized questions. For instance, in our studies we wanted to measure to what extent people would describe the YouTube channels that they watched as news channels. Based on their own viewing history, they were then asked to answer this question for their top 10 most watched channels, and had the option to select any news channels that came to mind from a searchable drop-down menu.

Data Donation flow

When participants open their application they are guided through a three-step process (Figure 1). In the first step, participants received instructions

[1]https://npm.io/package/data-donation-importers

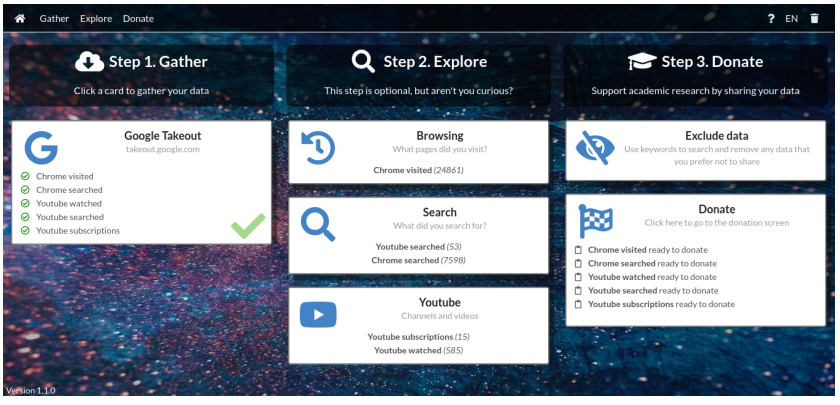

Figure 1: Home page of the data donation application after participants have performed the *Gather* step

for using the Google Takeout service to request their Browsing and Youtube history. We provided step-by-step instructions with pictures to make sure—insofar possible—that download options are set correctly. By only requesting the specific history data that we were interested in, the package size was generally below 2 Megabyte, and was often available for download within a minute after making the request.

In the second step, participants could explore their own data using an interactive wordcloud (e.g., top-visited web domains) and a table with the full data (e.g., URLs, time stamps). Users could search through the data using search strings, and delete anything they did not want to share. We explicitly marked this step as optional, which allowed us to investigate how many people would actually be interested in exploring their own data if they are not required to.

The final step was the donation. Users were first given an additional interface for filtering their data, in case they decided to skip the exploration step. Next, they we shown visualizations of their data and asked face validity questions. We also included an additional step for optional survey questions about the data, which we used in this study to ask questions about most frequently visited YouTube channels. Finally, participants were asked to donate their data.

Method

We conducted two data donation studies in which we asked participants to request their data from Google and donate it to us. Specifically, we asked

them to donate their Chrome browsing history, and Youtube viewing history. Chrome is the market leader for web browsers in the Netherlands, and browsing data was our primary interest in conducting these studies. Being able to see what specific content people visited online at what specific time can have great value for communication research, making the question whether people are actually willing to share this data with us all the more relevant. We wanted to included data from more platforms, but to keep the process simple we decided to only request a single DDP. We therefore decided on Chrome and Youtube, because both can be included in the same Google Takeout DDP.

Both studies followed a similar procedure. Participants first filled in a questionnaire that served to inform them about the data donation design, obtain informed consent, and collect demographic and attitudinal data. They were then directed towards the data donation application.

For the first study we recruited paid participants via a survey company that works with registered panel members. Next to a project management fee, we paid 44 euro (ex. VAT) per participant that completed the entire process, with participants expected to require 20 to 25 minutes. We paid a relatively high fee per participant to give them a higher monetary incentive than for a usual survey study, but the exact amount given to participants is not disclosed. To recruit participants the company contacted members via email, and provided a link to our Qualtrics questionnaire. Potentially interested participants could open the questionnaire to read the introduction to the study and see whether they pass the screening. The introduction explained that participants would first fill out the survey and then be directed to a separate application for donating their search and browsing history on Chrome and Youtube. Participants had to indicate that they understand this process and agree to participate, and through screening questions we excluded participants younger than 18, that do not have a Google account, or that never used YouTube.

Of the 9523 participants that opened the questionnaire, 3709 made it past the screening, 3652 completed the survey, and 435 donated their data. At each subsequent stage of drop-out we have additional information about the participants. For participants that dropped out during the survey we have basic demographics provided by the survey company. For participants that completed the survey we have several attitudinal measures. For participants that completed the survey but did not proceed with the data donation tool, the survey company inquired a random sample about their reasons. Finally, we also have log data for how participants used the data donation

application.

For the second study, we hosted a field lab on a large three-day music festival. At this festival there was a special research area where visitors could walk around and volunteer to participate in various studies. We were not allowed to recruit participants using a monetary incentive or physical gifts, and instead used the following three incentives. Firstly, participation should be educational and fun. Our slogan was "burst your own bubble" (referring to filter bubbles), and we explained that participants would be able to explore their own digital traces. We also kept a live scoreboard of the music preferences of all participants based on their YouTube history and a selected genre preference. Secondly, we emphasized that thee data would be used in actual scientific research and explained the importance of data donation for research. Thirdly, as a small gift that was sufficiently non-physical, participants could get a printed temporary tattoo for one out of four custom designs.

This time, participants in our study were assisted by lab personnel who would guide them through the process of gathering, exploring and donating the data. Likely in part due to this greater level of support, and the self-selection involved in visiting the special research area, almost all participants that showed interest in the study ended up donating (n = 326). Although this gives us less data to make inferences about predictors of participation, this study gave us valuable qualitative insight into how participants, and especially younger people, feel about data donation.

Face-validity of digital traces

There are many potential gaps in the data extracted from DDPs, and it is difficult to estimate how this affects our analyses. We therefore experimented with a simple method for measuring face-value validation. These self-reported measures have limited value as proof of validity, but as we demonstrate, they can definitely help to detect when data is flawed.

Participants were presented with visualizations of their data, and then had to answer a number of questions about this data. We used word cloud visualizations because these are easy to understand and can be applied to both label data (e.g., YouTube channels) and full-text (e.g., common words in search terms). In the survey study we asked four validation questions. We first asked a question about the ownership of the data: "Is the data only yours, or does someone else use your device or account?". Second, we asked about the accuracy of data: "Do you feel that you recognize this digital footprint as your own?". Finally we asked two additional accuracy questions that more

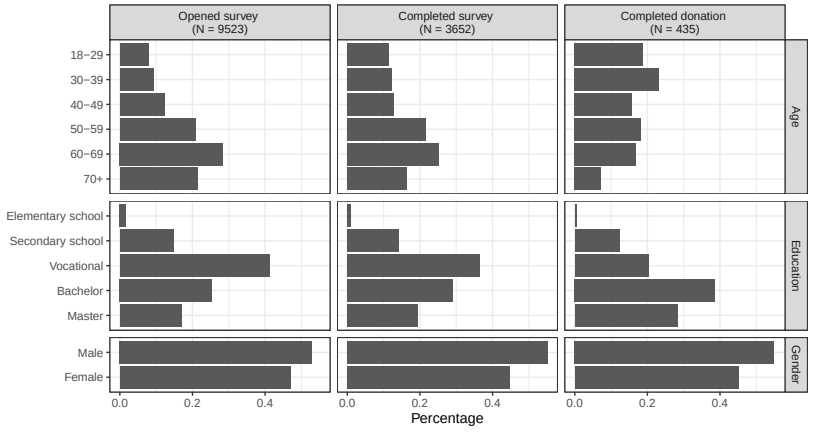

Figure 2: Participant demographics at different steps of the survey-based recruitment

specifically target *precision* and *recall*: "Are the largest items indeed the items you often visit?", and "Are there any items that you know you visited often, but are not shown here?". In the field study we only asked the first two questions, because the results from the survey study revealed that the additional accuracy questions provided mostly the same information.

Results

We first focus our attention on the survey-based recruitment. By analyzing the drop-out at different steps, we can see where we lose most participants and whether the drop-out is random or systematic. Starting at the point where participants clicked on the survey link sent to them by the survey company, we look at how many participants passed the informed consent and screening, finished the survey, and completed the donation.

The survey company provided demographic information of participants[2] even if they dropped out, which allows us to see how the drop-out affects demographic distributions. Figure 2 presents the distribution of age categories, highest completed education level and gender at each of the steps.[3] The biggest drop-out occurred during or at the start of the survey, where only 3709 of the initial 9523 participants agreed with the informed consent and made it past the screening. Of the participants that got passed this point

[2]For some participants the demographic information could not be provided.

[3]The panel company only more recently included other gender categories in the intake questions to panel members, and does not yet by default provide this information due to the re-identification risk

most also finished the survey (n = 3652). If we look at the shift in demographics, we see that the drop-out rate was slightly higher among participants that are older, female, and that did not finish a higher education degree. In particular, the proportion of participants older than 60 years decreased from 49.9% to 41.7%.

Of the 3652 participants that completed the survey, only 435 (11.2%) also completed the donation. In addition to losing many participants at this step, we also see a huge shift in the type of participants that continue. Drop-out was much higher among participants that are older and have a lower educational degree. The proportion of participants older than 60 years further dropped from 41.7% to 24%, and participants with a Vocational degree or lower dropped from 55.5% to 33.1%. To better understand why participants dropped out after finishing the survey but before donating their data, we conducted three additional types of analysis. First, we analysed predictors for what type of people dropped out, using attitudinal information obtained from the survey. Second, the survey company inquired from 100 random participants that dropped out at this point about their reasons. Third, we analyzed the log data of how participants used the application.

Predictors of dropping out after finishing the survey

We performed a logistic regression analysis to look more rigorously at what factors predict whether a participant that finished the survey also completed the donation. Next to the demographic information, we included two substantive measures from the survey. One is a measure for a persons general trust in other people, measured on a scale of four 7-point items. The other is a self-reported 10-point position on the political left-right dimension. We included these questions because a common concern about data donation is that certain parts of the population will be more difficult to reach, thus harming representativeness. Trust and political ideology in particular are relevant to the research we conducted using this data, and data biased along these dimensions can in general be harmful for external validity in communication research.

We included as a control variable whether participants finished the survey in under 3 minutes, because these participants are likely to have skipped the introduction. We also controlled for age in years, and included education as a 5-point ordinal scale. For gender we used a dummy variable for males. To control for age we used age in years, and for education we included the five categories as an ordinal variable. After removing cases with missing values, we have 3359 cases of which 432 donated their data.

Table 1: Logistic regression predicting which respondents that finished the survey opened and completed the donation (N=3359)

	Donated		
	base model	*model 1*	*model 2*
Predictors *(Odds Ratios)*			
Intercept	0.15***	0.28***	0.21***
survey < 3 minutes		0.56***	0.60**
Age		0.96***	0.96***
Male		1.50***	1.58***
Education		1.36***	1.27***
General trust			1.21***
Political left-right			0.95*
Deviance	2577.9	2416.2	2394.7
$\chi^2(df)$		161.74(4)	21.53(2)***
R^2 Tjur	0.000	0.050	0.057

*p<0.05; **p<0.01; ***p<0.001

Model 1 in Table 1 again shows the effect of demographic characteristics on participation, and the results reflect the results seen in Figure 2. However, when controlled for age and education, we now do see a substantial effect of gender, with the odds for males to donate data being 1.5 times higher.[4] We do not want to over-interpret causal mechanisms, but this provides further evidence of the additional response bias of data donation studies on top of the bias already present in the survey recruitment. Another notable observation here is that the 12.1% of participants that finished the survey in under 3 minutes were far less likely to complete the donation. Our interpretation is that there were quite some participants that skipped or rushed the informed consent page in the survey that explained the data donation component of the study. We discuss more evidence of this below.

In Model 2 we see that participants that score higher on the general trust scale are indeed more likely to donate their data. For every unit increase in trust (7-point scale) the odds increase by a factor of 1.21 (95% CI [1.09, 1.35]). If we look at the extremes of the scale, our model predicts that participants with the lowest trust score have a 6% probability of donating (95% CI [0.05, 0.09]) compared to 18% (95% CI [0.14, 0.23]) for the highest score. In our data trust

[4]The effect indeed disappears entirely when age and education are not controlled for.

was normally distributed (M = 4.000, SD = 1.071), and these extremes rarely occurred, with 71.03% of observations falling between 3 and 5, inclusive. We also find a negative effect, albeit very weak, of the political left-right dimension. For every unit that people move from the left towards the right, the odds of donating decrease by a factor of 0.95 (95% CI [0.91 - 1.00]). Both forms of response bias do not seem particularly harmful, but note that this is on top of any bias already present in survey participation. Moreover, this does indicate a limitation, or at least challenge, of using data donation methods for studying low trust or far-right communities.

Self-reported reasons for not donating

To better understand why participants dropped out after finishing the survey but before donating their data, the survey company inquired from 100 random participants about their reasons. Although only 22 replied to this request, the responses revealed a relevant mixture of three main reasons. The answers were given in Dutch and translated by the authors of this study.

The first reason is that many participants that did finish the survey appeared to still be unaware of the donation component of the study. Examples of answers that illustrate this include:

- "Don't want to upload."
- "Don't want to share my browsing history."
- "I'm not uploading my data, that is private and they do not need to know that."
- "I find this too personal, and too little insight into what will happen with my data".

The introduction to the survey clearly stated that participants would be asked to upload their personal data, and also required participants to indicate their informed consent before they could continue with the survey. Accordingly, these answers suggest that a number of participants did not fully understand what they signed up for. Even when asked to explicitly indicate informed consent, it seems that many do not read informed consent forms closely.

Of the participants that were not directly opposed to the idea of donating their data, the two main reasons for dropping-out were the amount of effort and the technical challenge. Despite our best efforts to make the process as simple as possible, the instructions for downloading the Google Takeout

data in the correct format are quite detailed, and participants need to follow the instructions closely.

- "Had the idea that I was far from finished and made the trade-off. Am I really going through all that trouble for a small reward?"
- "Too much effort."
- "Too much work and can't make it work."

The amount of effort is related to the technical challenge that participant experience, but the distinction is important. Where the participants that complained about effort might have been convinced by increasing incentives, there are also participants that seem to get stuck despite best efforts. The obstacle of the technical challenge could be the main reason for the relatively high drop-out of participants that are older and/or have a lower level of education.

- "It became to complicated for me. My understanding of computers is limited."
- "Become to hard for me and I get increasingly negative ideas about what would be shared. The Google page suggested this was a lot."
- "I stopped because I could not find the files."

Log data of the data donation application

At different steps in the data donation application we collected log data. Due to technical reasons we only obtained log data for approximately 75.63% of the participants[5], but this is enough to study some broader patterns in user behavior.

The first observation is that it seems that many participants that finished the survey (3652) did not proceed to the application (1596). If we estimate based on the percentage of missing log data that 2210 participants used the application, then around 1442 participants (31.3%) dropped out before even attempting to download their Google data. It thus seems that for these participants the problem was not in our specific data donation flow, but the fact that there was a data donation component to the study at all. Based on the self-reported reasons for not donating as discussed above, this can either be because they were unaware that they were expected to donate their data, or because they decided it would be too much effort.

[5]We estimated this percentage based on participants that did donate their data, but did not have log files

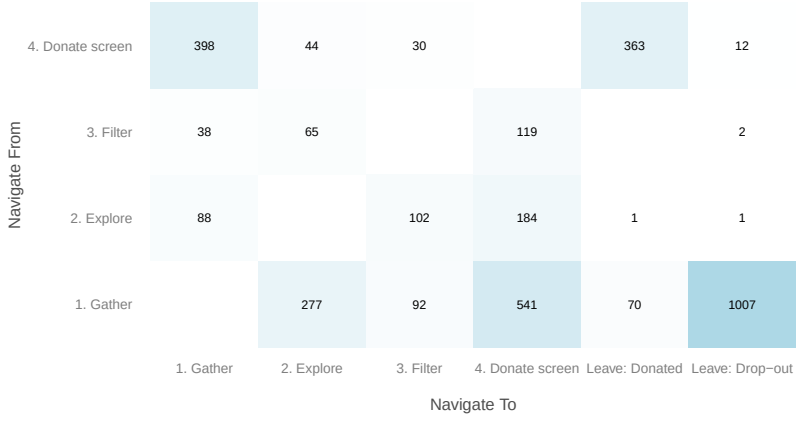

Figure 3: Application log data for survey study, showing the number of times participants navigated between sections, and from where they leave with or without donating.

Figure 3 shows how often participants navigated from one section (rows) to another (columns), and when they leave the application. We also documented when participants *Leave* the application, and based on whether they donated their data at that point we count this as either *Leave: Donated* or *Leave: Drop-out*. The most important observation is that by vast majority the people that dropped out did so during the *Gather* step, in which they had to download the data from Google. This is the most time-consuming and complicated part, and also the part that makes it very tangible and clear that we expect people to donate private data. When you follow the instructions, you first need to sign-in or re-authenticate at Google, which often involves 2-factor authentication. Once signed in, you need to manually tell Google what information you want to download, and then wait for Google to make the link available. This not only emphasizes that you're requesting sensitive data, but also means that participants need to be able to navigate several technical barriers, which previous research also identified as a critical bottleneck (Ohme et al., 2021; Struminskaya et al., 2021) in data donation research.

Another relevant observation from the log data is that in the survey study the explore page was hardly used. It seems that many participants were not very interested in seeing how much data Google actually has about their browsing and YouTube history, or in exploring their own digital footprint. It is plausible that this is also a results of the complicated process of downloading their Google Takeout data. After finishing the survey and downloading the

data, many participants might have simply wanted to finish the study.

Field study results

After the high drop-out rate in the field study, we had low to modest expectations about how successfully our effort to recruit participants at a music festival would be. While using lab personnel would remove the technical barriers and "hassle", the process would still take at least around 20 minutes, and there was no precedent to infer whether this is something that festival goers would sign-up for.

We were thus positively surprised to find out that visitors at the festival were very eager to participate in our study, and that most of them indeed found the experience interesting and fun. Of the 349 participants that agreed to participate and filled in the questionnaire, 326 participants completed the entire process. The actual number of people that showed interest was even higher, but this was the maximum number of participants that we could manage given the number of computers and lab personnel that we had available. There was often a line, and we actually had to send participants away because the waiting time would become too long. For the 23 participants that dropped out, a primary reason was technical issues with retrieving the Google Takeout data, which required logging in with 2-factor authentication.

There was also very little drop-out in the recruitment stage. Most visitors that came up to our lab to hear about the process decided to participate after hearing about the data donation aspect. Note, however, that there is a good dose of self-selection involved. It is safe to assume that people that visited the science area came with the intention to participate in a study. Visitors could also consult a flyer that briefly explained our study, so some visitors that visited our lab were already informed beforehand. Participants in the field study were also relatively young ($M = 28.05$, $SD = 6.79$) and higher educated (72.5% had a Bachelor or Master degree) compared to respondents in the survey study, where we also found that younger and higher educated people are more likely to donate their data.

The low drop-out rate in the field study leaves us short of empirical data to model what type of people would drop-out. But we did acquire a great deal of qualitative insight from the experience of assisting participants with the application, and from an open question regarding how they felt about donating their data. Overall, most participants reported that they did not at all mind sharing their data with us, and were very open about sharing and discussing their history data with the lab assistants. Interesting to note here

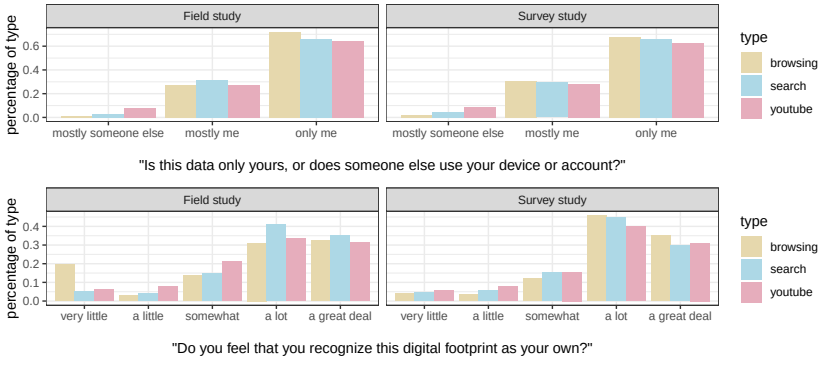

Figure 4: Self-reported accuracy and ownership of the donated digital traces in the field study (left) and survey study (right)

is that many participants also actively shared and discussed their data with their friends. Overall, they were much more open about their data than we expected.

Some participants emphasized the importance of the face-to-face context, saying that they would probably have not participated if alone behind a computer. This suggests that the field study context helps to alleviate not only technical barriers, but also trust barriers. For one, because a website is easier to fake than an official field lab at a renowned festival. But also because the face-to-face context made it much easier for us to convey why data donation is important for research, and to explain how we would use this data. If participants had any concerns they could ask questions.

One particularly interesting reason, voiced in different formulations, is that participants already "made peace with Google knowing everything", and that they were glad that this data could now also be used for science. We believe that this strongly conveys the value and legitimacy of data donation as a research method. Legislation like the GDPR empowers people to take hold of their own data and decide what to use it for, and data donation research can help people realize this. It enables citizens and scientists to collaborate towards the end of making sense of today's complex media landscape.

Validation questions

Two validation questions were asked in both the survey and field study. The results are presented in Figure 4. Results indicate that for most participants

the data was indeed mostly their own, or only their own. Still, these results also indicate that for a good percentage of participants the data also represents someone else, such as a partner, family member or friend. In addition to causing noise and bias in the analysis of these participants, this is also concerning from an ethical point of view. It means that we might be analyzing digital traces that belong to someone that did not give informed consent, and it is debatable whether it matters that this data is "owned" by the participant. Based on this validation question it would be possible to remove any data where ownership is uncertain, but at a hefty price of around 30% of the sample. Also, note that the answer "mostly someone else" is mostly given for YouTube data. In the field study, we observed that this is often the case for parents that let their children view YouTube on their accounts.

Figure 4 also shows that participants overall felt that the data gave an accurate depiction of their browsing, search and YouTube history. This was more so the case for the survey study. In particular, notice that the accuracy of the browsing data in the field study was very low for almost 20% of participants, whereas this was only the case for a few participants in the survey study. A plausible explanation is the strict screening in the survey study, where participants had to indicate that they use Chrome and YouTube, and that they could log in to their Google account. In the field study we did not reject any people from participating. It was quite common that festival visitors did not (consistently) use Chrome, but were still interested in exploring their YouTube history.

Figure 5 shows the results for two accuracy questions that were formulated to evoke a self-reported measure of precision and recall. We only asked these questions in the survey study, because the results correlated so strongly with the general accuracy question that we cut them out in the field lab to reduce completion time. The most important observation from these results is that the self-reported accuracy does seem to reflect both precision and recall. Of course, it cannot be taken for granted that this is always the case, and this is only a preliminary observation.

Reflection on Application design

To inform future tool development, we elaborate on some key considerations and lessons learned in designing and deploying our application.

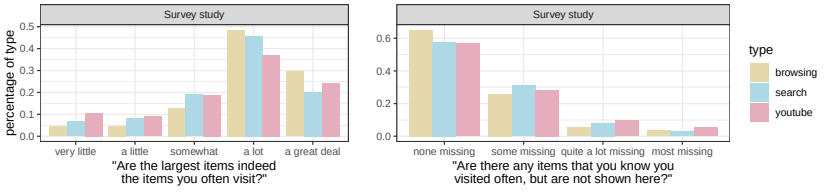

Figure 5: Self-reported precision and recall in the survey study

Limitations of strictly client-side processing

One of our design goals was to perform all the data processing client side (i.e., on the participant's own computer), so that participants could explore and filter the data before it ever touches our server. We still think this is a good design principle, but we also encountered two notable limitations.

The first limitation is that you can only let participants explore data that is explicitly available inside the Data Download Package (DDP), and sometimes the data that you want participants to see is not included. For example, at the time of writing the *TikTok* DDP only contains numeric IDs for the videos that people watched, without any textual information regarding the source and content of the video. To obtain this information we would need to scrape TikTok or use their API, but (unless the API is public) this requires using a server. This means that the data would have to leave the participant's device before they could have explored and filtered it. By not allowing this, we limit the possibilities for users to explore and filter data from certain DDPs.

The second limitation has to do with security. By (temporarily) storing data in the browser, we create an additional location on the computer where the data is stored, and could possibly be viewed by others that have access to the device. Developers also need to be careful to prevent unwarranted outside access to this data through cross-site scripting attacks. So while from an ethical and legal point of view it might seem better if the participant can explore and filter the data before giving consent and sending the data, from a security point the data can better be protected on a server.

Taken together, these two limitations provide reasons to reconsider the pros and cons of strictly client side processing. Our recommendation remains that client-side processing should be the default, but if there are ways in which server-side processing can enhance critical issues like informed consent and security, we should not dismiss them out of principle.

DDP import scripts

As explained in the *Application Design* section, we implemented a system for creating simple DDP import *recipes*. Through using this system, we observed that most DDPs can indeed be parsed with a simple set of instructions, thus avoiding the need for researchers to write custom import scripts in Javascript.

However, it is always possible that researcher encounter a DDP that cannot be parsed with such a system, so there should be an option to fall back to a programming language when more flexibility is needed. Perhaps ideally, this could be a hybrid solution, where common steps in the pipeline for importing and parsing DDPs are handled by a standardized instructions, and custom scripts are only used where needed. With the advancement of WebAssembly, these custom scripts could even be written in popular data science languages, such as Python and R[6] to run in the browser. For example, the recently developed *Port* application uses *Pyodide* to allow researchers to write the import scripts in Python (Boeschoten et al., 2023).

Data exploration and filtering features

In our application we explicitly made it optional for participants to use a dashboard for exploring, visualizing and filtering their data. This resulted in some observations and reflections about when and how such as dashboard can be relevant for a data donation application.

One purpose is to enhance informed consent. It facilitates participants to see and control what they are donating. From the field lab we know that some people really appreciate this, and even though it was hardly used in the survey study, we think that providing the option in itself satisfies an ethical purpose. However, we also learned that it can create the wrong expectations from users. At the field lab, a person that was knowledgeable about IT and privacy remarked how difficult it was to remove personal identifiable information (PII), since search history can reveal a lot about a person's living area, work, etc. What we failed to communicate properly is that it is not, and should never be, the responsibility of the participant to anonymize their own data. Dealing properly and securely with PII is the responsibility of the researcher.

Another purpose of a dashboard can be as an incentive or participation. At the field lab, we presented participation as an opportunity to explore your

[6]For more information, see the Pyodide (https://pyodide.org/) and WebR (https://github.com/r-wasm/webr) projects, for Python and R, respectively

own digital traces. When prompted this way, participants were indeed eager to explore their own data, and many enjoyed spending time with the tool. A question that remains is whether we could also have used this incentive better in the survey study. With regard to tool design, it could be effective to implement a preview of what the data exploration dashboard looks like before participants are asked to gather and import their own data.

In the current version of the tool we only supported filtering the data by search strings, and visualizing it as a word cloud. We recommend future initiatives to explore additional features and alternative types of visualizations. Based on how participants at the field lab liked sharing and discussing their data, we think there is potential in adding the option to download visualizations or share them on social media. We also had a screen that showed (non sensitive) aggregated data from all participants, and a similar feature that uses aggregate data could be integrated in the dashboard.

There are also some pitfalls that we encountered when implementing additional visualizations. Initially, we also included time charts to show Chrome history over time, or YouTube activity per hour of the day. We eventually dropped these because we encountered some complications, and were not able to fix and pilot them before conducting our study. The first is that the Google Takeout DDP presented the date in a locale (i.e. language) sensitive format, showing the names of months instead of numbers, and the locale used in the DDP was based on the user's settings. So when implementing time based visualizations, a tool developer needs to be careful to study the date formats used in the DDP, and/or use advanced date parsers. A second pitfall is that presenting multiple interactive visualizations side-by-side can be computationally demanding, since all the calculations are performed on the user's own device, which can be an old computer or mobile phone. Developers therefore need to think carefully about limiting CPU and memory use, and test on less powerful devices before deploying the application to a large group or participants.

Keeping it simple

Our application was designed to support DDPs from multiple platforms at once (e.g., Google, Meta, TikTok). In the application (Figure 1) there are then multiple cards in the *Gather* column, and the data produced by these DDPs is combined into categories in the *Explore* column. There are some benefits to this design, but it comes at the cost of simplicity. It is possible that one of the reasons that many survey participants dropped out shortly after opening the app, is that they felt overwhelmed and uncertain about how long the

process would take. If only one DDP is used, as in the current study, then it would have been possible to create a more simple and linear data donation flow. Based on the technical barriers with a single DDP, we also fear that working with multiple DDPs is simply not feasible in a survey study. For data donation studies with survey-based recruitment, we therefore think that the benefits of supporting multiple DDPs does not outweigh these limitations. The current design of our application seems more suitable for a field lab context, in which the participant is assisted by lab personnel that knows how to use the tool.

Conclusion and discussion

This paper explored the potential of data donation as a method for collecting participant-centered digital trace data. We presented two studies in which participants were asked to request their data from Google and donate it to us using a newly developed web application. For one study we recruited participants with a classic online panel survey, and for the other study we organized a field lab on a large music festival. We collected data specifically for the purpose of reflecting on successes and challenges in recruitment, data validity and the design of the application.

As expected based on prior research (Reiss et al., 2022), the drop-out rate during the survey study was very high. Of the 9523 participants that opened the link to our study, only 3709 agreed with the informed consent and passed the screening. Among the 3652 participants who completed the survey, only 435 finished the donation. Our findings also verify concerns of response bias in data donation, on top of any response bias in survey participation. People that donated were on average younger, higher educated and male. They also scored higher on general trust scores, and more often self identified as politically left-wing. Overall, this indicates that even within a survey sample, that is likely already biased in some ways, there are certain characteristics that make participation in data donation research less likely. There are likely to be more factors that we need to take into account, such as technical knowledge and privacy attitudes. If data donation participants are recruited via surveys, we recommend designing the survey so that substantive variables of interest (e.g., data privacy attitudes, technical knowledge) are asked before screening out participants based on willingness to donate. This allows one to see and possibly correct for sample bias, and also contribute to data donation methodology by identifying what factors correlate with willingness and ability to donate.

Participants that dropped out in the data donation stage stated as rea-

sons that they did not want to donate their data, or considered the process to take too much time and effort. This indicates that many of them did not fully read or understand the informed consent form at the start of the survey, where they indicated willingness to donate their data. This is problematic, because the informed consent form also serves to explain the process and why we ask them to donate, and this information could potentially improve participant retention. A recommendation for future research is therefore to use emphasis techniques for informed consent forms (see e.g. Varnhagen et al., 2005) to highlight the data donation component at the beginning of the study.

Of the participants that did have the intention to make the effort to donate their data, there was still much drop out due to technical barriers. This is a known bottleneck in data donation research (Ohme et al., 2021; Struminskaya et al., 2021), and based on our log data analysis we can pinpoint that for our study the main bottleneck occurred in the process of requesting and downloading the Data Download package (DDP). This makes it a difficult problem to solve, because many DDPs can only be obtained directly from a specific website (e.g., takeout.google.com), meaning that we cannot integrate this in our applications to make the process easier. The potential of data donation research is therefore directly related to how companies are willing to provide the data. For the field of data donation research to move forward, a difficult but important avenue is to engage with companies and political actors, and contribute to maturing the field of researching with data rights (Ausloos & Veale, 2020).

One of most exciting outcomes of our field lab study is that it showcased how much of a difference it makes to conduct data donation research in a face-to-face setting. Participant stated that this made them trust the method more, and by taking away the technical barriers the process was much faster and enjoyable. The field study also proved much more effective at fulfilling the three goals of data donation research: collecting data, involving participants in the data collection process, and enhancing their understanding of their own digital footprint. In the survey study the first goal was achieved, but the second and third goal were left mostly untouched. In the field study, we experienced firsthand that participants could indeed show great interest in their data, and that it could help them put into perspective that their digital footprints are indeed being recorded. It made many participants more aware of what the GDPR is, and why this type of legislation is important for themselves as well as for researchers. Creating this type of awareness and positive experiences is not just a boon for the participants themselves, but

could also move the field forward by creating awareness and familiarity with the concept of data donation. It is safe to assume that most of our survey respondents had never heard of data donation, and that this affects how comfortable they are with donating their own data.

Overall, we would thus recommend researchers that are considering a data donation study to also consider the option of using a lab setting where participants are assisted by humans. Survey-based recruitment can in potential be relatively cheap and better for obtaining a representative sample, but depending on the type of data and complexity of the data donation flow drop-out can be very high and systematically exclude certain types of people. Without a massive budget, recruiting a large and representative sample might simply not be feasible. In a lab setting, the costs per participant are generally higher, but in return it offers great opportunities for more in-depth and qualitative means of analysis. In terms of methodological contributions, the opportunity to sit down with participants and evaluate their data together can also yield valuable insight into how people feel about donating data and how valid they believe the data is.

Our results for the validation questions reveal that participants generally felt that their data gave an accurate depiction of their browsing, search, and YouTube history. More importantly, we demonstrated that with two simple questions we can at least identify participants for whom the data is *not* accurate at all. The question about data ownership in particular provides very relevant limitations to consider, as we found that around 25% to 30% indicated that their data was not entirely but *mostly* their own, and a few participants had data on their account that belonged to *mostly someone else*. This presents a real risk for the quality of data donations, and researchers should especially take this into account when requesting data for platform that are commonly shared by multiple users (e.g., Netflix, News applications) Furthermore, it raises ethical concerns regarding whether or not this data can be used, and highlights the need for better ways to determine ownership.

There are notable limitations of using questions to validate the donated data that should be taken into account, and can be improved on. Most importantly, if participants indicate that they do not recognize a digital footprint as their own, we do now know whether this is because the data is inaccurate, or because of the participant's memory and self-image. The answers will also be affected by the formulation of the questions and the visualization of the digital traces, so more research is needed to establish best practises. To compare approaches and determine their usefulness for measuring data quality, one could for instance conduct experiments where

the digital footprint is manipulated by adding or removing data. Based on our experience in the field lab, we also believe that more qualitative research where the researcher can go back and forth with the participant can shed more light on the validity of data donation data.

The application that we developed and used for these studies is made available open-source.[7] We aim to maintain it and make it as easy as possible to deploy. But more importantly, we hope that our experiences with developing this iteration will be of use for future tool development, in the same way that our application was inspired by WebHistorian (Menchen-Trevino, 2016) and OSD2F (Araujo et al., 2022). Notwithstanding that our current tool works well for what we set out to do, there are still many factors to concern for making data donation software as effective and secure as possible, and different types of DDPs and studies will also require different designs. Some improvement can be built into the current tool, but others will benefit from having a fresh start. To prevent re-inventing the successful parts of the wheel, we should strive to make these parts available as modular components. From our current application, we therefore published the pipeline for parsing a DDP based on a set of standardized recipes as an open-source NPM module.[8]

We conclude that data donation is a very promising method, but noting that it is uncertain whether and how fast some of the core promises will be fulfilled. A clear avenue ahead is to keep working on and experimenting with different tools and recruitment strategies, and to systematically investigate their efficacy. The more misty road is that the technical barriers—that seem to be a critical bottleneck—can not always be solved in our study designs, but are entwined with data rights, the whims of big companies, and public perception towards the act of donating data. A risk is that data donation research will be railroaded by the digital traces that we can (more easily) collect, similar to how the study of social media content has often been railroaded by API access to the major platforms. This is not a reason for avoiding data donation, but it calls for researchers to think critically about how it fits into their toolkit, and when to best use it.

A final thing to consider is that data donation is a new method that most of all still needs to mature, not only in how we use it as researchers, but also in the public eye. We were genuinely surprised by the level of enthusiasm and interest that we experienced in hosting the field lab. When given the opportunity to properly explain why we seek to involve people so closely

[7] https://github.com/ccs-amsterdam/DigitalFootprintsLab
[8] https://npm.io/package/data-donation-importers

in the data collection process, a vast majority showed great willingness to get involved. Perhaps one of the big steps forward towards improving participant retention, is simply to spread the good word.

Acknowledgements

The field lab would not have been possible without the help from our dedicated team of lab assistants: Puck Guldemond, Judith Stevens, Cathérine Minczeles, Niels Kuipers, Teun Siebers, Alexandra Schwinges, Nike Soffree, Kinga Ławicka & Anne Kroon.

This research was funded by multiple grants: the Dutch Research Council (NWO) under both a VENI (VI.Veni.191S.097) and JEDS (Inside the Filter Bubble, 2017) grant; the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 947695);

References

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., van de Velde, B., De Vreese, C., & Welbers, K. (2022). Osd2f: An open-source data donation framework. *Computational Communication Research*, *4*(2), 372–387.

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, *11*(3), 173–190.

Ausloos, J., & Veale, M. (2020). Researching with data rights. *Amsterdam Law School Research Paper*, (2020-30).

Boeschoten, L., de Schipper, N. C., Mendrik, A. M., van der Veen, E., Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, *8*(90), 5596. https://doi.org/10.21105/joss.05596

Bowser, A., Hansen, D., He, Y., Boston, C., Reid, M., Gunnell, L., & Preece, J. (2013). Using gamification to inspire new citizen science volunteers. *Proceedings of the first international conference on gameful design, research, and applications*, 18–25.

Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2022). User-centric approaches for collecting facebook data in the 'post-api age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, 1–20.

Christner, C., Urman, A., Adam, S., & Maier, M. (2022). Automated tracking approaches for studying online media use: A critical review and recommendations. *Communication methods and measures*, *16*(2), 79–95.

Dvir-Gvirsman, S. (2017). Media audience homophily: Partisan websites, audience identity and polarization processes. *New media & society*, *19*(7), 1072–1091.

Facebook. (2013). *Hadoop*. https://www.ReactJS.org

Janic, M., Wijbenga, J. P., & Veugen, T. (2013). Transparency enhancing tools (tets): An overview. *2013 Third Workshop on Socio-Technical Aspects in Security and Trust*, 18–25.

Maus, B., Salvi, D., & Olsson, C. M. (2020). Enhancing citizens trust in technologies for data donation in clinical research: Validation of a design prototype. *Companion Proceedings of the 10th International Conference on the Internet of Things*, 1–8.

Menchen-Trevino, E. (2016). Web historian: Enabling multi-method and independent research with real-world web browsing history data. *IConference 2016 Proceedings*.

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital trace data collection for social media effects research: Apis, data donation, and (screen) tracking. *Communication Methods and Measures*, 1–18.

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the ios screen time function. *Mobile Media & Communication*, *9*(2), 293–313.

Pfiffner, N., & Friemel, T. N. (2023). Leveraging data donations for communication research: Exploring drivers behind the willingness to donate. *Communication Methods and Measures*, 1–23. https://doi.org/10.1080/19312458.2023.2176474

Prior, M. (2013). The challenge of measuring media exposure: Reply to dilliplane, goldman, and mutz. *Political Communication*, *30*(4), 620–634.

Reiss, M., Pfiffner, N., Mitova, E., & Blassnig, S. (2022). Strategies to collecting digital trace data through data donations for communication research. *Paper presented at the ECREA conference*.

Rossi, A., & Lenzini, G. (2020). Transparency by design in data-informed research: A collection of information design patterns. *Computer Law & Security Review*, *37*, 105402.

Schufrin, M., Reynolds, S. L., Kuijper, A., & Kohlhammer, J. (2020). A visualization interface to improve the transparency of collected personal data on the internet. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1840–1849.

Shadbolt, N. (2013). Midata: Towards a personal information revolution. *Digital enlightenment yearbook*, 202–224.

Skarlatidou, A., Ponti, M., Sprinks, J., Nold, C., Haklay, M., & Kanjo, E. (2019). User experience of digital technologies in citizen science. *Journal of Science Communication*, *18*, E. https://doi.org/10.22323/2.18010501

Skatova, A., & Goulding, J. (2019). Psychology of personal data donation. *PloS one*, *14*(11), e0224240.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field.

Struminskaya, B., Lugtig, P., Toepoel, V., Schouten, B., Giesen, D., & Dolmans, R. (2021). Sharing data collected with smartphone sensors: Willingness, participation, and nonparticipation bias. *Public opinion quarterly*, *85*(S1), 423–462.

van Atteveldt, W., & Peng, T.-Q. (2021). *Computational methods for communication science*. Routledge.

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures*, 1–17.

Varnhagen, C. K., Gushta, M., Daniels, J., Peters, T. C., Parmar, N., Law, D., Hirsch, R., Sadler Takach, B., & Johnson, T. (2005). How informed is online informed consent? *Ethics & Behavior*, *15*(1), 37–48.

Verbeij, T., Pouwels, J. L., Beyens, I., & Valkenburg, P. M. (2021). The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports*, *3*, 100090.

Veys, S., Stamos, M., Reitinger, N., Mazurek, M. L., & Ur, B. (2020). Designing visualization and exploration tools for data access under gdpr/ccpa.