Beyond data transactions

A framework for meaningfully informed data donation

Gomez Ortega, Alejandra; Bourgeois, Jacky; Hutiri, Wiebke Toussaint; Kortuem, Gerd

DOI
10.1007/s00146-023-01755-5

Publication date
2023

Document Version
Final published version

Published in
AI and Society

Citation (APA)
Gomez Ortega, A., Bourgeois, J., Hutiri, W. T., & Kortuem, G. (2023). Beyond data transactions: A framework for meaningfully informed data donation. *AI and Society*, *40 (2025)*(2), 405-422. https://doi.org/10.1007/s00146-023-01755-5

AI & SOCIETY (2025) 40:405–422
https://doi.org/10.1007/s00146-023-01755-5

MAIN PAPER

Beyond data transactions: a framework for meaningfully informed data donation

Alejandra Gomez Ortega[1] · Jacky Bourgeois[1] · Wiebke Toussaint Hutiri[2] · Gerd Kortuem[1]

Received: 4 November 2022 / Accepted: 16 August 2023 / Published online: 30 August 2023

Abstract
As we navigate physical (e.g., supermarket) and digital (e.g., social media) systems, we generate personal data about our behavior. Researchers and designers increasingly rely on this data and appeal to several approaches to collect it. One of these is data donation, which encourages people to voluntarily transfer their (personal) data collected by external parties to a specific cause. One of the central pillars of data donation is informed consent, meaning people should be *adequately informed* about what and how their data will be used. However, can we be *adequately informed* when it comes to donating our data when many times we don't even know it is being collected and, even more so, what exactly is being collected? In this paper, we investigate how to foster (personal) data literacy and increase donors' understanding of their data. We introduce a Research through Design approach where we define a data donation journey in the context of speech records, data collected by Google Assistant. Based on the data donation experiences of 22 donors, we propose a data donation framework that understands and approaches data donation as an encompassing process with mutual benefit for donors and researchers. Our framework supports a donation process that dynamically and iteratively engages donors in exploring and understanding their data and invites them to (re)evaluate and (re)assess their participation. Through this process, donors increase their data literacy and are empowered to give meaningfully informed consent.

Keywords Data donation · Voice assistants · Personal data · Data literacy

1 Introduction

Most people interact daily with products and services that generate, collect, and (indefinitely) store data about them (e.g., digital platforms, and smartphones). Data are valuable for researchers and designers across several fields, including healthcare and well-being (e.g., Choe et al. (2018); Low et al. (2020)), design (e.g., Bogers et al. (2016); Gorkovenko et al. (2019)), ubiquitous computing (e.g., Bourgeois et al. (2014); Tolmie et al. (2016)), and artificial intelligence (e.g., Martelaro et al. (2021); Liao and Sundar (2021)), It offers insights distributed over time, new forms of participation, and perspectives to engage with and prompt reflection. For this reason, researchers and designers rely on different ways to collect data. Data donation is an emerging approach to data collection. It is a voluntary transaction of (personal) data (Skatova et al. 2014); where people donate their data to researchers or designers who will use it in a specific context (e.g., Gómez Ortega et al. (2022); Razi et al. (2022); Breuer et al. (2022); Cooper et al. (2022)).

One of the central pillars and key challenges of data donation is informed consent (Ohme and Araujo 2022; Bietz et al. 2019; Jones 2019; Strotbaum et al. 2019). To provide informed consent, a person must be *adequately informed* and have a clear understanding on *what* and *how* her data will be used (Neisse et al. 2016). Previous research argues that in the context of Big *observed* data—implicitly

✉ Alejandra Gomez Ortega
a.gomezortega@tudelft.nl

Jacky Bourgeois
j.bourgeois@tudelft.nl

Wiebke Toussaint Hutiri
w.toussaint@tudelft.nl

Gerd Kortuem
g.w.kortuem@tudelft.nl

[1] Faculty of Industrial Design Engineering, Delft University of Technology, Landbergstraat 15, Delft 2628 CE, The Netherlands

[2] Faculty of Technology and Policy Management, Delft University of Technology, Jaffalaan 5, Delft 2628 BX, The Netherlands

created as a byproduct of people's actions (U.S. Chambers 2014; Bowyer et al. 2022), it is difficult for people to be *adequately* informed (O'Connor et al. 2017; Neisse et al. 2016; Andreotta et al. 2021).

One scenario that illustrates this difficulty is speech records: data collected and stored by voice assistants (e.g., Google Assistant, Siri, and Alexa). Speech records are *observed* data generated in the background of the (un) intended interactions between people and their devices. Voice assistants are always listening and activate when users use the wake word, *"OK Google*", *"Hey Siri*", or *"Alexa*". Afterward, they process, respond to the user's query, and store a *speech record*; containing a timestamp, transcript, and audio recording.[1] Previous research demonstrates that most voice assistant users have an incomplete understanding of how speech records are collected, stored, and processed (Lau et al. 2018; Pins et al. 2021), as well as the security and privacy implications (Chalhoub et al. 2021; Malkin et al. 2019). Hence, their understanding of the information and infrastructure behind speech records prevents them to be *adequately* informed when pondering whether to donate (or share) them with researchers.

In this paper, we introduce a data donation case study where we investigate how to foster (personal) data literacy and increase donors' understanding of their data. Our case study is grounded in the context of speech records, which are increasingly used by researchers (e.g., Pins et al. (2021); Bentley et al. (2018); Malkin et al. (2019)) and where supporting better-informed decisions is crucial. We hypothesize that fostering a better understanding of the data (i.e., increased (personal) data literacy) enables data donors to be *adequately informed* about and (re)evaluate their participation and yields a return of value for them. Specifically, we investigate the following research questions:

RQ1. How do donors describe their (personal) data literacy throughout the data donation process?
RQ2. What do donors perceive as the value they gained through data donation?

To address our research questions, we adopt a Research through Design (RtD) approach (Zimmerman et al. 2007; Giaccardi and Stappers 2017). We defined, designed, and developed a data donation journey providing a fully functional embodiment of data donation mediated by a digital platform. Throughout the journey, 22 Google Assistant users (i.e., donors) donated their speech records. They reflected on their understanding of the data and their data donation experience via short questionnaires and semi-structured interviews, which we analyzed through reflexive thematic analysis (Braun and Clarke 2006, 2013). Our findings indicate that supporting an incremental understanding of the data enables donors to be *adequately* informed. Additionally, the knowledge and empowerment derived from the increased understanding of the data and how it relates to and reflects a person's behavior are perceived as valuable.

We offer our recommendations in the form of a framework that conceives data donation as an integral process around voluntary transactions of data (as opposed to an instance where the transactions occur). We propose to situate informed consent in a broader set of activities before, during, and after the voluntary transactions of data. Data donors, as active participants, dynamically and iteratively engage in exploring and understanding their data and are invited to (re)evaluate and (re)assess their participation.

2 Background

2.1 Data donation

Data donation is an approach to data collection where people contribute to research by voluntarily sharing their personal data. It is defined by Skatova and Goulding (Skatova and Goulding 2019) as the act of a person *actively consenting* to donate, or transfer, their personal data to a specific cause (e.g., a scientific research project). Researchers across several fields, including philosophy of technology, healthcare, data journalism, design, and human-computer interaction, have engaged with the concept of data donation both conceptually and empirically. Conceptually, research has focused on understanding the characteristics of data donation as a transaction (e.g., Prainsack 2019a; Hummel et al. 2019) and identifying best practices for ethical data donation (e.g., Bietz et al. 2019; Krutzinna et al. 2019; Ohme and Araujo 2022). These include ensuring that: (1) data donors are *adequately informed* regarding their participation (Ohme and Araujo 2022; Bietz et al. 2019; Jones 2019; Strotbaum et al. 2019) and (2) the relationship between data donors and receivers is not *'starkly unbalanced'* (Prainsack 2019a); meaning data donors should *derive value* from engaging in data donation (Prainsack 2019a; Hummel et al. 2019; Bietz et al. 2019).

Empirically, research has focused on understanding donors' motivations (e.g., Skatova and Goulding 2019; Skatova et al. 2014; Diethei and Niess 2021) and (privacy) concerns (e.g., Rudnicka et al. 2019; Maus et al. 2020) and approaching, or developing ways to approach, (personal) data collection through data donation (e.g., Gómez Ortega et al. 2022; Razi et al. 2022; Breuer et al. 2022; Cooper et al. 2022). Broadly, previous research has relied on two ways for people to donate their data:

[1] In response to the European General Data Protection Regulation (GDPR) as of 2020, voice assistants only store the audio recordings if the user has opted in.

1. By installing a mobile app (e.g., Robert Koch Institut 2020) or web plug-in (e.g., Breuer et al. 2022; Malkin et al. 2019) that integrates with a third-party application and collects the data.
2. By requesting and downloading a copy of the data from a third-party application and uploading it to a dedicated platform (e.g., Gómez Ortega et al. 2022; Razi et al. 2022; Pins et al. 2021; Cooper et al. 2022).

In the second approach, individuals obtain (a copy of) their personal data and (re)use it (e.g., to contribute to scientific research). It is enabled by recent changes in privacy regulations, such as the General Data Protection Regulation (GDPR) (GDPR 2018) in Europe that proposes the *rights to access and data portability*.

Breuer et al. (2022) compared the two ways to donate personal data in the context of Facebook logs. They argue that it is possible to obtain informed consent from donors in both ways and the second one offers higher transparency to donors; who can *"see exactly what types of data they will share with the researchers"*. Yet, this argument fails to consider (1) donors' general (and limited) understanding of their data at the time of informed consent and (2) the practicalities of obtaining a copy of the data, which are not trivial. First, whether donors authorize integration with a third-party application or upload a copy of their data, being adequately informed entails understanding the content of the data and its (privacy) implications. It is often not the case (see Sect. 2.2), especially when it comes to *observed* data, already collected through data donation, such as speech records (Malkin et al. 2019; Pins et al. 2021), sensor data from fitness trackers (Cooper et al. 2022; Robert Koch Institut 2020), and Facebook (Breuer et al. 2022) or Instagram (Razi et al. 2022) logs. Previous research has partially addressed this challenge; for instance, in the context of menstrual tracking logs, Gómez Ortega et al. (2022) developed a tool for donors to explore and visualize their data *after* providing informed consent and manually delete it if necessary. Yet, at the time of informed consent, often prior to and independent of the data donation (e.g., Razi et al. 2022; Robert Koch Institut 2020; Breuer et al. 2022; Malkin et al. 2019), the content of the data remains obscure and abstract. Second, arguing that uploading a copy of the data offers higher transparency assumes donors can obtain information from it. Bowyer et al. (2022) conducted a study inviting 11 people to obtain a copy of their data from different organizations; they found that in most cases people were left 'in the dark' with files that were hard to understand and make sense of. Alizadeh et al. (2019) conducted a similar study and concluded that people require support in understanding and making sense of the files and the data. Hence, donors might be able to see the files of data they will share with the researchers but might not *adequately* understand the (personal and sensitive) information they contain (and that they are giving away).

Concluding: When collecting (personal) data through data donation it is fundamental to ensure that (1) data donors are *adequately informed* regarding their participation and (2) data donors *derive value* from engaging in data donation. Due to the nature of the data, most donors are not necessarily *adequately informed* regarding its content and (privacy) implications at the time of informed consent. We address this challenge by proposing a data donation journey centered around fostering (personal) data literacy. In doing so, we approach informed consent as a dynamic process and invite donors to (re)evaluate their participation as they better understand the content of the data. Furthermore, we hypothesize that a better understanding (i.e., increased (personal) data literacy) yields a return of value for donors.

2.2 Supporting personal data literacy

Data literacy is broadly related to the ability to understand the information that can be obtained from data (Wolff et al. 2017). It is often associated with a more or less specialized skill set according to a person's role and needs (Wolff et al. 2017; Clegg et al. 2020). For instance, a data scientist, a high school student, and a person interacting with digital technologies engage with data with different objectives and in distinct situations; they each require specific skills. We focus on the necessary skills of a person interacting with digital technologies that collect (personal) data[2]. In this context, data literacy has centered on inviting people to understand and question how data fits into their lives (Gray et al. 2018; Pins et al. 2021). For instance, Gray et al. (2018) propose the concept of *'data infrastructure literacy'* as the ability to *account for*, *intervene around* and *participate in* the wider socio-technical infrastructures through which data is created, stored and analysed.

Previous research demonstrates that most people need support to develop personal data literacy because (1) they are unaware of their rights concerning personal data management (Van den Berg and Van der Hof 2012; Bowyer et al. 2022) and (2) it involves data-intensive activities such as data exploration, interpretation, and sense-making (Pins et al. 2021; Kurze et al. 2020; Jakobi et al. 2018). Researchers from the fields of (usable) privacy (e.g., Jakobi et al. 2018; Tolmie et al. 2016; Kwon et al. 2018) and data visualization (e.g., Kurze et al. 2020; Pins et al. 2021; Pu et al. 2021) have explored several ways to support people in becoming aware of and understanding data collection and processing. For instance, Tolmie et al. (2016) developed a

[2] Defined in the GDPR as information related to an identified or identifiable person (GDPR 2018).

prototype that supported the legibility of sensor data (e.g., temperature, humidity, motion) collected at home and invited people to interpret and account for the data. They argue for supporting interpretation and *'articulation work'* to foster personal data legibility. These activities involve various orders of reasoning (e.g., place, time, people, practices, and events) and lead to relating data to specific events and reflecting through data. Similarly, Pins et al. (2021) developed a prototype that supported the exploration of voice assistant data (e.g., speech records) and invited people to interact with it. They provide design recommendations to foster awareness and support (personal) data literacy, including: (1) support right from the start (e.g., guiding people in obtaining a copy of the data), (2) support to structure the data into categories, (3) reconstruct the context, (4) draw attention to unintended interactions, (5) tell users how the vendors (might) see them, and (6) disclose the communications between devices and services (e.g., third-party services and applications).

Concluding: We approach (personal) data literacy in terms of the necessary understanding that a person must have to make *adequately* informed decisions about their data. We build upon the work and recommendations by Pins et al. (2021) and Tolmie et al. (2016) to propose a data donation journey centered around fostering (personal) data literacy. Specifically, we focus on supporting data exploration and understanding (e.g., cleaning, structuring, visualizing) as well as interpretation and articulation (e.g., reconstructing the context of the data).

2.3 Voice assistants

Voice assistants are routinely used by millions of people around the world as part of their daily and social lives (Pins et al. 2021). It is reported that in 2022 Google Assistant and Apple's Siri are each used by over 500 million people worldwide, while Amazon's Alexa is used by over 100 million people worldwide[3]. Users of voice assistants integrate these devices into various tasks and activities throughout the day, including managing smart appliances, getting ready for bed, and cooking (Bentley et al. 2018; Sciuto et al. 2018). Every time a user interacts with a voice assistant (e.g., 'OK Google, what is the weather like?'), the device generates and stores a *speech record*. Speech records correspond to *observed data* (U.S. Chambers 2014; Bowyer et al. 2022) as they are indirectly collected in the background of the (un) intended interactions between people and their always-on devices (Pins et al. 2021; Malkin et al. 2019). They contain (1) timestamp describing when the interaction occurs (date and time); (2) transcript describing the content of the interaction (what); and (3) audio recording describing who initiated the interaction and how (e.g., loud or quiet environment). Thus, speech records allow for a detailed picture of voice assistant users and their routine activities (Pins et al. 2021).

Previous research suggests that most voice assistant users have an incomplete understanding of how speech records are collected, stored, and processed (Lau et al. 2018; Pins et al. 2021), as well as the security and privacy implications (Chalhoub et al. 2021; Malkin et al. 2019). Bentley et al. (2018) collected speech records via Mechanical Turk from 88 users (Google Assistant), they concluded that users interact with these devices approximately between 2 and 18 times per day, with an average of 4.1 times per day. Hence, it is difficult for voice assistant users to be aware of what information is stored on their speech records, especially over time. Similarly, Pins et al. (2021) argue that it is difficult for voice assistant users to understand the extent of data collection and processing by the system or vendor; as these are introduced through vague and unclear terms of use statements and privacy policies. They developed a prototype to support exploration where 11 users (Alexa and Google Assistant) uploaded a copy of their speech records. Moreover, Malkin et al. (2019) developed a browser extension to retrieve speech records from 116 users (Alexa and Google Assistant); they used individual speech records as survey prompts and found that almost half of the users (51.7 %) did not know their speech records were permanently stored and the majority (56.0 %) did not know they could review their past speech records.

Concluding: Broadly, users are hardly informed about the data collection practices of voice assistants and the nuances of information collected and stored in every speech record. Yet, speech records are collected and used for research activities (e.g., Pins et al. 2021; Malkin et al. 2019; Bentley et al. 2018). Hence, speech records are a relevant context where (1) data is already being requested and shared and (2) most people are not adequately informed when consenting to share their data.

3 Methodology

In this paper, we investigate how to foster (personal) data literacy and increase donors' understanding of their data through data donation. In doing so, we aimed to enable data donors to (1) be *adequately informed* about and (re)evaluate their participation, and (2) *derive value* from engaging in data donation. In particular, we sought to understand:

RQ1. How do donors describe their (personal) data literacy throughout data donation?

[3] Voice assistant users worldwide, from smartspeakersglobalmarket-report (accessed in September 2022).

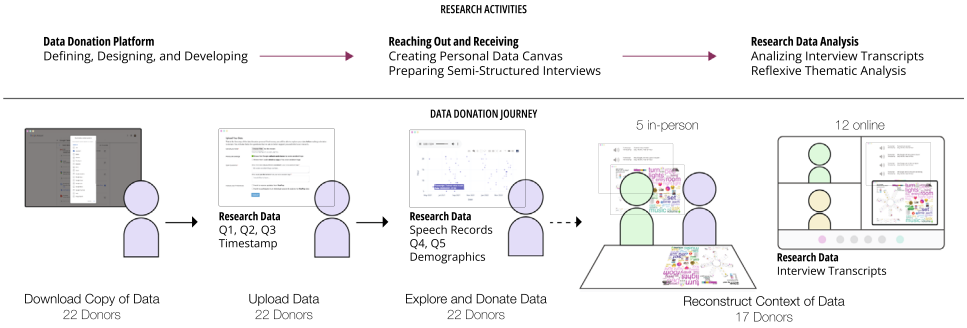


Fig. 1 Research activities and data donation journey

RQ2. What do donors perceive as the value gained through data donation?

To address these research questions, we adopted a Research through Design (RtD) approach (Fig. 1) (Zimmerman et al. 2007; Giaccardi and Stappers 2017). We defined, designed, and developed a data donation journey (Sect. 3.1) and platform (Sect. 3.2) to collect speech records generated by Google Assistant—a context where most people are not adequately informed, yet, data is being requested and used for research. 22 Google Assistant users (i.e., donors) (Sect. 3.3) donated their speech records and reflected on their understanding of the data and their data donation experience via short questionnaires and semi-structured interviews that we analyze using reflexive thematic analysis (Sect. 3.4). Our institution's Human Research Ethics Committee and Privacy Team reviewed and approved these activities.

3.1 Data donation journey

The data donation journey (Fig. 1) aimed to provide a concrete embodiment of data donation, mediated by a platform, grounded in the theoretical principles proposed by Gómez Ortega et al. (2022). Specifically *awareness*, as an opportunity to foster (personal) data literacy and increase donors' understanding of their data. We approached data donation by inviting potential donors to request and download a copy of their data from a third-party application and upload it to a dedicated platform, similar to previous research (Gómez Ortega et al. 2022; Razi et al. 2022; Pins et al. 2021; Cooper et al. 2022). Throughout the journey, we incorporated the recommendations by Pins et al. (2021) and Tolmie et al. (2016) by setting a space to support interpretation and articulation of the donated data; and we established explicit instances for donors to (re)evaluate their choices. In doing so, we embodied data donation (and informed consent) as an encompassing process (i.e., data donation journey) as opposed to an instance (i.e., when a person donates her data).

3.1.1 Downloading the data

The first step in the data donation journey was for donors to download a copy of their data from Google. Here, we aimed to foster (personal) data literacy by highlighting that (1) speech records are collected and stored by Google and (2) users can obtain a copy of their speech records (and other data). Following the recommendation by Pins et al. (2021) of "supporting right from the start" we provided detailed visual instructions describing the process. Donors were required to:

(1) visit takeout.google.com[4] and log in with their Google credentials;
(2) select the type (i.e., speech records), format (i.e., JSON, and MPEG) and size of the data to export;
(3) wait *'a long time (possibly hours or days)'*[5] for the export to complete; and (4) receive an email with a ZIP file containing their speech records. Here, it is important to note that donors receive via email a ZIP file containing a JSON file listing all speech records and several MP3 files (one per speech record). As described by previous research (e.g., Bowyer et al. 2022; Alizadeh et al. 2019) these files and formats are hard to understand for most people.

3.1.2 Uploading the data

After receiving an email containing a copy of their speech records, donors could upload their ZIP file to the data donation platform. When uploading their data to the platform, donors could find information about the research project (i.e., project goals and activities) and team (i.e., names, affiliations, and contact information of the researchers) and detailed visuals describing the data management and storage.

[4] Google Takeout Page: takeout.google.com.

[5] Once people complete the Google Takeout process they see the following message: "Google is creating a copy of files from My Activity. This process can take a long time (possibly hours or days) to complete. You'll receive an email when your export is done."

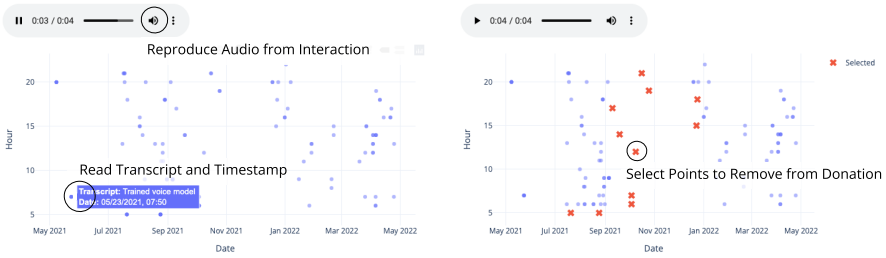


Fig. 2 Example of interactive graph where potential donors can explore an overview of their speech records over time

(a) Hovering over points to explore. (b) Selecting points to delete.

With this activity, we aimed to (1) *adequately* inform donors about our research goals and activities and (2) support them in (re)evaluating their participation. When uploading the data, donors provide informed consent to participate in the research. But they have not (yet) consented to donate (transfer) their speech records to the researchers. Hence, data is uploaded and stored in the platform, but researchers do not have access to it until donors have explored it and explicitly (re)evaluated their decision. Additionally, in this step, we collected data on donors' initial understanding of Google's data collection and storage practices via a short questionnaire. We invited donors to answer the following questions[6]:

Q1. Did you know that Google collects and stores your speech records?
Q2. Did you know that you could download a copy of your Google Assistant speech records?
Q3. What information do you think is in your speech records?

Answers from this questionnaire served as a baseline to determine donors' initial awareness with respect to that of participants in previous studies (e.g., Lau et al. 2018; Malkin et al. 2019; Chalhoub et al. 2021).

3.1.3 Exploring, understanding, and donating the data

After uploading their data, donors were invited to explore it, understand it, and (re)evaluate their participation. Here, donors could assess whether and what data to donate (transfer) to the researchers. With this activity, we aimed to foster (personal) data literacy by enabling donors to delve into the (1) content (2) dimensions (i.e., timestamp, transcript, audio recording), (3) amount, and (4) temporal distribution of the uploaded speech records.

To support and enable exploration, we build upon the data visualization prototype developed by Pins et al. (2021), who visualize speech records as points on a graph arranged by time to help people reason about it. We augment their prototype by allowing donors to *listen* to their audio recordings when hovering at a point, in addition to reading the transcripts and timestamps. Hence, donors could visualize (and listen to) an overview of their speech records over time through an interactive graph where each point represents an interaction with the Google Assistant (Fig. 2). In the graph, the *x*-axis represents the time of the day and the *y*-axis represents the date. When donors hover over a point, they can listen to the audio recording and read the transcript and the exact date and time of the interaction. Together with the visualization, donors could see the following message: *"we invite you to explore **(and listen to!)** your data by hovering over the dots, each dot represents an interaction with your voice assistant."*

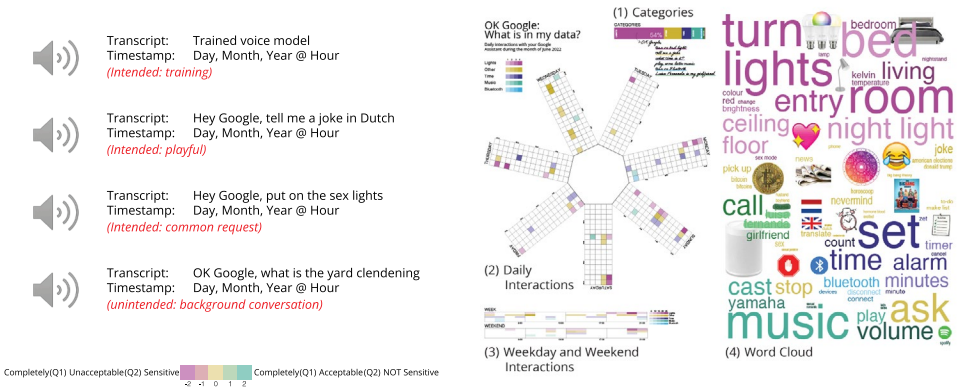
After exploring their data, donors were invited to (re) evaluate their participation by choosing to: (1) withdraw from participating and delete all their speech records from the platform; (2) consent to donate (transfer) all their speech records to the researchers, who immediately gain access to it; (3) remove specific speech records (e.g., single data point, all data from a given time) from the platform and consent to donate (transfer) the remaining speech records to the researchers. If donors consented to donate their speech records, we invited them to provide their demographics (i.e., self-described gender, age), location (i.e., city), and information about their Google Assistant (i.e., device type, language). Additionally, in this step, we collected data on donors' understanding (after the exploration) of their speech records via a short questionnaire. We invited donors to answer the following questions:[7]

Q1. To what extent do you agree with the following statements?

(a) Seeing my data is helpful for understanding how much data my assistant collects
(b) Seeing my data is helpful for understanding what types of data my assistant collects

[6] Answers to Q1, Q2 are Yes/No, answers to Q3 are open.

[7] Answers to Q4 are 5-Point Likert scale from strongly disagree to strongly agree, answers to Q5 are open.



(a) Single interactions. (b) Multiple interactions.

Fig. 3 Example of a personal data canvas. Shown with permission of the donor

(c) Seeing my data is helpful for understanding how long my assistant has been collecting data
(d) Seeing my data is helpful for deciding whether to donate them

Q2. What did you learn from seeing your data in this way?

Answers from this questionnaire served to determine donors' understanding of their data after the exploration and their perceptions regarding the usefulness of the visualization.

3.1.4 Reconstructing the context of the data

The last step in the data donation journey was for donors to reconstruct the context of their data. This step is informed by previous literature pointing out a need to support contextualization, interpretation, and articulation (e.g., Tolmie et al. 2016; Pins et al. 2021; Gómez Ortega et al. 2022). Participation in this step was voluntary, and only donors who opted in were invited to participate. Here, we aimed to foster (personal) data literacy by facilitating donors' exploration and interpretation of their speech records. To support exploration and interpretation, we developed a *personal data canvas* following the recommendations by Pins et al. (2021) for voice assistants' data literacy: (1) "drawing attention to unintended interactions", (2) "supporting to structure the data into categories", and (3) "telling users how the vendors (might) see them".

The *personal data canvas* introduces speech records as single and multiple interactions. First, (Fig. 3a), we focused on introducing the dimensions of the data (i.e., timestamp, transcript, audio recording) through single interactions and "drawing attention to unintended interactions". Second, (Fig. 3b), we focused on introducing multiple interactions through a data visualization. In doing so, we supported donors to "structure the data into categories", and broadly "told users how Google (might) see them". In the visualization (Fig. 3b), we focused on conveying the information from the timestamps and transcripts of multiple interactions. Specifically, we identified common interactions for each dataset and grouped them into categories (e.g., weather, music, time). We visualized the distribution of these categories throughout the dataset with a bar graph (Fig. 3b$_{(1)}$), and we represented each category with a different color throughout the visualization. Additionally, we presented the number of (daily) interactions for each category per hour of the day and day of the week with a heat map (Fig. 3b$_{(2)}$) where we focused on the 16 h of the day with more interactions, the start and end times vary by the donor. Similarly, we used a heat map to present the number of interactions of each category per hour of the day during the weekdays (Monday through Friday) and weekends (Saturday and Sunday) (Fig. 3b$_{(3)}$). Finally, we presented a word cloud (Fig. 3b$_{(4)}$) with the most frequent words grouped and color-coded by category, and additional images were visually representing some of the terms. We added the images to make the interactions more prominent and easier to explore.

We used the *personal data canvas* as a prompt during semi-structured interviews where we supported donors in exploring the data, reflecting on their behavior (as captured by the data), and identifying patterns and potential inferences. This approach has been successful in previous research (e.g., Bogers et al. 2016; Bourgeois et al. 2014; Kurze et al. 2020; Malkin et al. 2019). From the interviews, we collected qualitative data on donors' understanding of their speech records and overall data donation experience. The interviews revolved around three stages. First, we invited donors to describe their data donation experience (up to that point, *downloading*, *uploading*, and *exploring, understanding, and donating* the data) and whether they considered removing points from their donation. Second, we introduced the *personal data canvas* and explored the different attributes of the data in terms of sharing and sensitivity. Here, we supported donors to lead and articulate

the interpretation and contextualization of their data. Third, we invited donors to describe any feelings or emotions that emerged throughout the data donation experience (comprising the interview) and discuss their perspectives on the value gained. We phrased value gain in terms of getting something out of the experience or wishing something for a future experience[8]. During the interview, we reminded donors of the possibility of (re)evaluating their participation and withdrawing their donation.

3.2 Data donation platform

We designed and developed a digital (web) platform[9] for people to donate their speech records to our research. Existing platforms, such as Open Humans[10], allow the sharing of data collected by third parties but these do not yet allow the exploration and granular selection of specific types (and points) of data to share, which was critical to foster (personal) data literacy, increase donors' understanding of their data, and support donors in (re)evaluating their participation and setting boundaries. On the platform, donors could find information about the research project (i.e., project goals and activities) and team (i.e., names, affiliation, and contact information of the researchers) and detailed visual instructions describing the process of downloading a copy of their data and the data donation journey. In addition, they could explore their data as described above (Sect. 3.1.4) and (re) define the terms and boundaries of their participation (e.g., whether and what data to donate, whether to participate in reconstructing the context of the data, delete all data). Once donors upload their data into the platform, it is stored on a database that is only accessible to them (and the system administrator). If they consent to donating their data, it becomes accessible to the researchers. At all times, donors can revoke their consent (i.e., data is no longer accessible to the researchers) and delete their data from the platform (i.e., data is no longer stored on a database).

The platform has three open source components that manage (1) the user profiles and authentication, (2) the data storage and sharing, and (3) the donation process. The first two were implemented using TypeScript, and the third was implemented using the Python web framework Django. Data was passed between system components using web APIs.

3.3 Participants: data donors

Between April and June of 2022, we reached out to Google Assistant users worldwide (e.g., Assistant App, Google Home, Google Nest) and invited them to participate in our research by donating and reconstructing the context of their data. For this, we used a combination of convenience and snowball sampling. We advertised our research by periodically posting on our personal social media (e.g., Twitter, LinkedIn), existing online communities (e.g., subreddit r/googleassistant, local mailing lists and newsletters), posting flyers in local cafes and universities, and advertising our research at community events.

Twenty-two donors, aged 21–58 years (mean = 30.8, median = 38), 1 identified as non-binary, 7 identified as female, and 15 identified as male, positively responded to our call by donating their data. Donors were primarily located in the Netherlands (54%), with some based in other countries, including Germany, Italy, Colombia, and Argentina. Obtaining a copy of the speech records, enabled by the GDPR, was also possible for donors outside the EU[11]. 17 donors (5 identified as female and 12 as male) agreed to participate in a follow-up data exploration interview. The first author conducted the interviews in English between June and July 2022. Interviews lasted between 35 and 55 minutes; 5 took place in person and 12 via Zoom. The *personal data canvas* was presented as two slides on a screen; if the interviews were in person, the visualization (Fig. 3b) was also printed on A3 paper. We conducted one interview with the two members of a household who share a device ($D9_{a,b}$), the remaining 16 interviews were one-on-one as most donors were single-users of their Google Assistant. The interviews were audio recorded and transcribed. The first author made an initial transcript using MS Office 365, then manually reviewed and edited it.

3.4 Data analysis

Throughout the data donation journey, we collected the following data: (1) speech records, (2) demographics, (3) answers to questionnaires (Q1–Q5), and (5) interview transcripts. Our analysis primarily focuses on the answers to the questionnaires (Q1–Q5) and interview transcripts. The answers to the closed questions (Q1, Q2, Q4) are used to illustrate donors' understanding of their data at different points of the data donation journey. The answers to the open questions (Q3 and Q5) are combined with the interview

[8] Example of questions: What did you get out of this experience? What would you like from a data donation experience in the future?

[9] Data donation platform and open-source code can be accessed at: https://datadonation.ide.tudelft.nl/

[10] Open Humans: https://www.openhumans.org/.

[11] The GDPR applies to the population of the European Union. Yet, in practice, the right to data portability is available worldwide, since international companies rarely limit it by geography (Bowyer et al. 2022).

data, and analyzed by the first two authors using reflexive thematic analysis (Braun and Clarke 2006, 2013), within a constructionist framework. Both authors independently read through the transcripts to familiarize themselves with the data and coded the entire dataset using ATLAS.ti. Through this process, we aimed to capture all the aspects of the data relevant to the data donation experience and the perceived value gained. Both authors independently reviewed the codes and subsequently discussed and grouped them into tentative themes. The first author iteratively reviewed and refined the themes.

4 Results

In this section, we first describe the themes illustrating donors depictions of their (personal) data literacy and perceived value gain throughout the data donation journey. Then we introduce donors' perspectives on their data donation experience and illustrate some of the difficulties they encountered as well as our shortcomings (Sect. 4.5).

4.1 Getting my data

Donors gained awareness of Google's data collection practices and the possibility of obtaining a copy of their data. Over half of the donors (12 out of 22) indicated not knowing it was possible to obtain a copy of their data **(Q2)**. *"I didn't know, when I saw in the beginning like the instructions about how to download this data. It was the first time for me, and actually, it was very interesting"* (D16). For them, reading the *call to donate* and the *instructions on how do donate* was a way to discover their rights and with them new ways of engaging with their personal data. Additionally, throughout the download process, data went from an abstract entity to a nearly tangible (and material) one that is available and can be explored digitally, seen, read through, and listened to. Data is there, stored somewhere, and accessible (to donors and others).

4.1.1 Data literacy

Donors' (personal) data literacy increases by understanding how to intervene and participate in Google's data collection practices and becoming familiar with their individual (data) rights (e.g., right to data portability). This understanding extends beyond the context and scope of our research, as described by $D9_b$.

> "What I found the most interesting was, while I was downloading the data, to see how organized it was. We were only following instructions, so we deselected everything and then we uploaded just the voice commands. But I was genuinely excited to see that I could look up my YouTube history, my Google searches, my Google Maps. Everything is in a specific folder, so I think this research empowers you to look stuff up that you otherwise wouldn't look for" ($D9_b$)

It illustrates how all kinds of (personal) data are stored in structured databases and how these are searchable and accessible upon request. Hence, it presents the opportunity for donors to access and explore data from other Google services (e.g., browsing history, location) and other data holders (e.g., Spotify, Twitter) if only out of curiosity.

4.1.2 Value gain

Donors engage differently with their (personal) data; an abstract concept that gained clarity and materiality. Speech records are opaquely generated as a product of the (many) interactions between people and their voice assistants and are stored *"somewhere in a cloud"* (D19). They became available and inspectable through the act of 'obtaining a copy'. Moreover, speech records became something donors *have*, *own*, *control*.

> "Can you say that you own a dataset or that the data about your life is yours if you are not really capable of using it, or donating it, or doing anything about it? Because, I feel like that data [the speech records] is Google's data. I mean, if I don't have a server, if I don't have the technical ability, if I never use it in my daily life, is that data mine? It's about me, but I don't really feel it is mine. Thanks to this project, we kind of gain ownership over that. If I'm a passive agent, I feel like it is about me, but if I'm an active agent it is mine" ($D9_a$)

$D9_a$ articulates the difference between data being *about her* and *hers*. Through data donation donors gained ownership of data that is *about them*; and became *theirs*. Here ownership is not limited to *having* a copy of the data. It extends to **actively** being able to *control*, *guard*, and *use* it; for example, by deciding to donate it. Nonetheless, *having* is important. It means donors can (re)use the data as they wish; although donors acknowledge this process is not straightforward and it requires technical skills and resources.

4.2 Exploring leads to knowing

Through inspecting and exploring the data donors become aware of *"how much information is stored and the kind of information that is stored"* (D14); and how data relates to themselves and their interactions with Google Assistant. Exploration led to the paradoxical realization that interactions are recorded and become data, paradoxical given that most donors (17/22) indicated being aware that Google

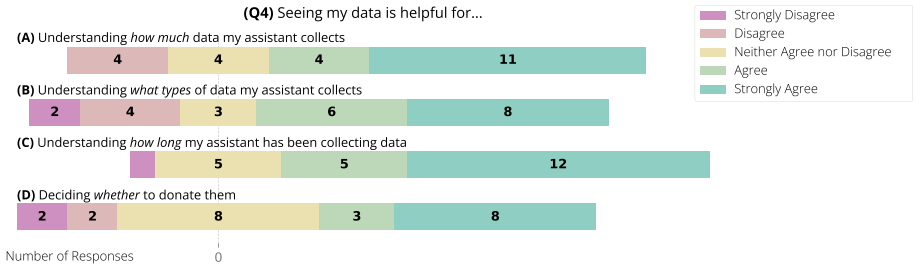


Fig. 4 Responses to **(Q4)**, indicating donors perceptions of the data exploration tool

collects and stores their speech records when answering to **(Q1)**. It underlines the gap between using a device *knowing* that it collects and stores (personal) data and *knowing* how data *looks*, *sounds*, and *feels like*.

> "I don't know how to describe it. But one thing for me is like to use it [Google Assistant] and the other thing is like listening to my voice now. By listening to it, you become aware of the fact that this is recorded and was stored somewhere by Google, and it makes you feel a bit unsettled" (D5).

4.2.1 Data Literacy

Donors' (personal) data literacy increases by realizing that the interactions with Google resulted in data points, indefinitely stored *"Google records all this information for, I don't know how long"* (D18). This realization was described as *surprising* (D2,D14,D17,D21), *crazy* (D14), and *creepy* (D1, D19). Especially considering the *observed* nature of speech records, collected implicitly and in the background of all interactions donors have with their Google Assistant. It is nearly impossible for a person to keep track of every interaction over time, *"I cannot remember what kind of stupid things I've said to Google"* (D18); hence *knowing* how data *looks*, *sounds*, and *feels like* can lead to uncomfortable feelings, *"Google feels like a stalker"* (D19), and emotions.

> "I remember when I downloaded it, I hadn't made a Google takeout before. I did not know that they stored all this data. And like all the data that is there, that for me was a moment of real emotional response like oh, OK" (D21)

The data visualization supported donors in *knowing* how data *looks*, *sounds* and *feels like*. In addition, it helped most donors grasp how much data was collected, what types of data were collected, and for how long (Fig. 4). Further, it enabled donors to identify behavioral patterns in their data, *"there are patterns on the time of the day when I'm using the assistant, which reflect, somehow, my routine* (D22). In this way, donors gained awareness of data being personal; related to themselves and their behavior and reflecting specific aspects about themselves and their behavior.

4.2.2 Value gain

The gained awareness and familiarity, often described as *knowledge* and (increased) *understanding*, were perceived as valuable takeaways from participating in the data donation journey. These led to a (better) informed opinion on the data and how it relates to themselves and others.

> "What I took out of this experience is knowledge. It's knowledge about what Google collect[s], the fact that you can download the data, listening to ourselves, it was exciting. And then, having an informed opinion about [data]. Before I had like a fear, and now I have an informed opinion, or at least semi informed" ($D9_a$)

Donors appreciated how data donation enabled them to look 'behind the curtain' and gain knowledge into how Google 'sees' them and how it works; what it listens to, *"listening to it, it's like oh OK, it was recorded"* (D5); what gets recorded, even when it should not, *"it was a personal conversation, I was not aware that* [it] *was being recorded"* (D14); what it does (and does not) understand, *"the transcripts are not always the things that I said"* (D18); and how much and how often it collects and stores data, *"I could realize how much information is on my phone, about me"* (D16).

4.3 Knowingly giving away my data and contributing to research

In addition to supporting donors' understanding and exploration of their data, the data visualization enabled them to be aware of what exactly they were 'giving away' to our research, *"it is interesting to hear the recordings because you get a sense that there is a level of control of what you're giving away"* ($D9_b$). Here, data became a way for donors to participate in and support research activities; informed and enabled by increased (personal) data literacy.

> *"We were able to listen to the recordings and it's very intimate, but it kind of gave me peace of mind because at the end of the day, what is there is not what I value the most when I think about my privacy"* ($D9_a$)

4.3.1 Data literacy

Donors' (personal) data literacy enables them to be aware of the information donated (or transferred) to researchers. This awareness, invites them to reflect on (and (re)evaluate, if necessary) their privacy boundaries. We aimed to further support donors define their boundaries and *control* what they were 'giving away' by enabling them to remove specific points from their donation. Yet, most of the donors (21/22) decided **not** to remove any points, "I didn't find anything that was recorded that I thought well, no, no, I don't want to share it" (D10). This was primarily motivated by how most data points corresponded to simple and mundane interactions (e.g., 'OK Google, what time is it?', 'OK Google, set an alarm'), *"when I looked at the data, it was really, like, 'How is the weather', so, I did not see anything that I would have needed to remove"* (D2).

4.3.2 Value gain

Through the data donation journey donors *knowingly and actively* contributed to our research. Donors expressed having high regard for scientific research and considered our research *a good cause*, *"when I donate, also money I wanted to go for a good cause. And I'm convinced, it makes me convinced, that my data went for a good cause"* (D2). Hence, the action of contributing to research was perceived valuable and led to positive feelings. The motivation to contribute to research also shaped donors' decisions on whether to donate their data *"I understand the research process and I understand that they* [researchers] *need this kind of information, so I'm completely open to do it"* (D17), and which data to donate, *"I feel that if I share more data, more interactions, it will be more useful for the research. So, my decision was to help as much as I can"* (D14). Additionally, having *knowingly and actively* contributed to our research meant donors had expectations regarding the research progress and its outcomes. These underline opportunities for researchers further provide value to donors by being open and accountable.

> "If I took the time to donate my data to a project, it's because I'm actually interested in it. So, I want to keep knowing what is happening or how is my data being used for" (D1)

4.4 Reflecting on my data, my relationship with Google, and myself

During the semi-structured interviews donors gained deeper insights into how (personal) data relates and reflects aspects about themselves and their behavior. These, however, are incomplete and limited by the specific ways in which people interact with their Google Assistant, *"it does give a sort of accurate picture, but it's not the picture that I would put together. I think it gives, let's say to certain topics, more, uh, prominence than how prominent they are"* (D2). For example, the interaction 'OK Google, turn off the alarm' can indicate when a person wakes up, and the interaction 'OK Google, turn off the bedroom lights' can indicate when a person goes to sleep; broadly reflecting her sleep routine but not providing any insight into what happens in between. Yet, although limited and incomplete, data can support reflection. D22 illustrates the process of reflecting on her routine through the (lack of) data:

> "Realizing about Fridays, that I don't use Google on Fridays. It's like why? And then I thought like yeah, OK so I was not at home on Fridays. I mean, my working day ended like at noon. So, it was like super interesting, I didn't realize about those patterns [before]. Because I know that I wake up at 5:30 and I go to sleep at 10, but then I was never, like, OK the last three Fridays I was doing this" (D22)

4.4.1 Data literacy

Donors' (personal) data literacy increases by engaging with the data and reflecting on the nuances of the context captured by it. Through this process, donors realized the many ways data is embedded in and partly reflects their daily lives and interests and the potential inferences that could derive from data. In doing so, they gained greater awareness resulting in a *tipping point* in their perspectives on (personal) data. Data was no longer considered *nothing* (i.e., simple and mundane) and became *something* (i.e., personal and sensitive).

> "When I signed up for this study, I was like, OK, my Google home data? I don't think there is anything to find in it, so why wouldn't I share [it]? And even after ticking and ticking through [in the data visualization on the data donation platform], like OK, what am I sharing? I was still convinced. [...] And now I'm surprised, it's not like there is nothing in the data. For a brief moment, I was even like, OK, I'm glad that nothing more surprising came out from there [laughs]" (D2)

Increased awareness, and the change of perspective derived from it, resulted in the intention to change how donors interact with their Google Assistant to minimize (sensitive) data collection, *"I learned about myself, but also, I think I would be a little bit more careful with what I'm going to ask Google from now"* (D17). Additionally, it enabled donors to put their privacy concerns into perspective. For some, it led to the realization that Google is not that bad, *"I mean it might be able to tell if I'm sleeping, or at what time do I wake up, but like the things that are really important for me in*

terms of privacy are not there and so I was kind of relieved" ($D9_a$). For others, it led to the realization that Google collects and stores too much data and is too creepy, *"Google feels like a stalker"* (D19). It prompted donors to reconsider their relationships with digital technologies as well as their participation in our research and future research activities that entail personal data sharing. Still, when invited to (re) evaluate their decision to donate at the end of the data donation journey none of the donors (0/22) wanted to withdraw their donation.

> "It was a learning process for me. To understand better what information I share and what information I let Google know about me. And it's something that I believe we need to improve, because as I told you, there are things in the [personal data canvas] that you did that I don't want to be sharing with anyone, even with Google" (D14)

4.4.2 Value gain

Donors appreciated the multiple viewpoints set in place for them to explore their data. These allowed them to engage with data through different lenses and direct their attention to specific details, including amount, temporal distribution, (un)intendedness, aggregation, and potential inferences. The knowledge gained during the process was incremental. It led donors to challenge their assumptions (e.g., my data is just 'How is the weather') and account for the personal nature of the data; that relates to and reflects aspects of their behavior, especially when combined and considered over time.

> "Even though the data, if you take 1 by 1, is not something important or relevant. Those behavioral patterns are quite sensitive, like what you did [with the personal data canvas], like you can infer what my days look like a little bit" (D1)

The incremental knowledge donors gained throughout the process was enabled and supported by the guidance and materials we provided. These were highly appreciated and perceived as valuable takeaways in themselves. Especially the *personal data canvas* (Fig. 3b), *"it's a very nice visualization. Specifically, I appreciate the visualization"* (D17). It provided a structure to interpret the data, *"this distribution* [Fig. $3b_{(1)}$], *I just love it"* (D21); and prompted donors to reflect on their behavior.

> "It is super interesting seeing all this data like classified, as it was. [$3b_{(2)}$] And as I said, it helped like identifying the patterns of my day, of my routine" (D22).

The *personal data canvas* became a tangible outcome of the process, which we gave to the donors at the end of the interview, *"for me, getting back this visualization is useful. It really makes sense and it can actually tell something about me"* (D2). Beyond the material, donors found value in the guided exploration of their data. This process offers the opportunity to bring to light (personal) insights and disentangle the abstract construct of (personal) data. Hence, it could be relevant even beyond the context of our research.

> "I would pay for this. I would pay to have this kind of consultancy. Like not having to go into Amazon, and Spotify, and Facebook, and Twitter to understand. I think I would be one of the people that would pay for someone to go, explore and tell me from the platform's perspective how is my life. Kind of like people who pay for astrology? For other people to tell you who you are, you are fearless, you are... [laughs] Definitely, if I had someone doing data explorations with me, that would be something I would be interested in paying [for]" ($D9_a$)

4.5 Reflections on the data donation experience

4.5.1 A new experience

Data donation is a new experience. It enables donors to engage with their personal data and in doing so, open their personal space to others (i.e., researchers). This process can be confronting and uncomfortable. In the context of *speech records*, which correspond to *observed data* that is generated implicitly from people's behavior, this process is also a window to the unknown and the unexpected that is entangled in and captured by the data.

> "I must say that the I feel a bit ***naked***. In the sense that this [personal data canvas] tells a lot about me, much more than I expected" (D17)

Additionally, data donation entails to 'give away' a copy of their personal data. Although this is something donors do *knowingly and actively*, it is a leap of faith. Meaning, donors 'give away' their data to researchers in a specific context and under certain conditions. But they cannot guarantee that researchers will use their data in said context and under said conditions. They can only trust.

> "I'm relieved there's not more data out there. And that just triggered also thoughts in me about, like, how this data donation is really cool, but I'm also giving you permission to like do whatever with it" (D2)

4.5.2 An (not so) easy process

Data donation, as operationalized in this research, entails a journey that comprises several steps and interacting with at least three digital platforms (i.e., Google Takeout, email provider, data donation platform). This was by-design as

we aimed to support and promote awareness throughout the process. Yet, it meant that some donors faced difficulties and were confused.

> "It's not exactly super confusing, but it's confusing enough that I feel like I'm not sure if I'm doing this right when I'm downloading the data, when I'm uploading the data, I was usually expecting something different to happen" ($D9_b$)

The complexity of the journey might discourage potential donors, especially those less experienced in interacting with digital technologies. Hence, there is a need to balance the awareness gained throughout the process with its complexity.

4.5.3 A window into-the-wild, or not?

On 2020 Google announced that the Google Assistant was no longer collecting and storing the audio recordings from every interaction unless users had explicitly opted-in to allow voice data collection. We were aware of this and when disseminating the *call to donate* we invited potential donors to check their configuration and, if necessary, opt-in to allow voice data collection and interact with their Google Assistant for a couple of weeks or months before donating their data. Four donors (4/22) opted-in to allow voice data collection and generated data while being aware that it was going to be used for our research, *"when we turned it on, we were like wow,* [Alejandra, first author's name] *is going to listen to this"* ($D9_a$). We instructed these four donors to interact naturally with their Google Assistant. Yet, being aware of our research led to interesting interactions (e.g., 'OK Google, what do you know about [Alejandra, first author's full name]') and behaviors.

> "It was also interesting because I had the settings on and I had some guests. So first, I thought well, I have to make a little note [saying] that you can be recorded. Then I forgot, and after the visit I thought maybe I have to inform them. So, I did informed them afterwards." (D10)

Previous research claims that the data that is available through data donation 'is embedded into the donors' routine and is not attached to a research project or a research instrument, thus [is] less prone to observation bias' Gómez Ortega et al. (2022). Our research, where this partially applies (18/22), illustrates how data donation is limited by the infrastructures in which data is embedded in.

4.5.4 An (not so) individual journey

> "If I'm trying to donate data to some project, then I then I would like to donate data that I know it has picked up from me, and well..." (D2)

In this project we received 22 datasets and we identified more than one speaker being recorded in all of them (22/22). Different speakers were more frequent in multi-user environments (e.g., $D9_{a,b}$), where more than one person shares a physical space where a device is present. Yet, there were still recorded in single-user environments where other people (e.g., occasional visitors) are around (e.g., D2). Hence, although donations were made by a person *knowingly and actively* giving away her data- –except for $D9_{a,b}$ who gave away *their* data—other people were indirectly involved and information about them (e.g., their voice) was donated.

Donated data captures people's relationships and interactions with others (e.g., partners, family members, friends, neighbours), who are present in the dataset. Moreover, it accounts for people's relationships with others (e.g., 'OK Google, call *my mom*', 'OK Google, *my girlfriend* is [name]'). Hence, data that is donated could indirectly involve other people, who are captured by the data, *"I thought, well, it is my uncle's privacy, I don't want to compromise, someone else's privacy"* (D10). Underlining the importance of accounting for them. Although doing it is not necessarily trivial. Donors expressed having informed others (e.g., partners, family members, friends) of the data collection and the data donation, *"I would want my partners OK that this data is being shared"* (D11). But in some cases, figuring out who to inform can become a puzzle.

> "Oh, first of all, it's not even me. [laughs] I don't think I know who [it is]. 10th of December. Because, it [the Google Assistant] is close to a window, so, but I don't think the window was just open. Like stuff from the street" (D2)

5 Discussion

5.1 (Adequately) informed consent

In this study, we led donors on a journey of engaging with and understanding their (personal) data. We focused on speech records, collected by Google Assistant in the background of users' every interaction. An underlying outcome of this data donation journey, was illustrating the gap between (1) *knowing* that voice assistants collect and store data; (2) *knowing* what data *looks*, *sounds*, and *feels* like; (3) and *knowing* how it relates to a person and reflects her behavior. This gap challenges the notion of being *adequately* informed in situations that involve a transaction of *observed* personal data. It echoes the limitations of informed consent highlighted in previous literature (O'Connor et al. 2017; Neisse et al. 2016; Andreotta et al. 2021; Brown et al. 2016).

Breuer et al. (2022) argue that it is possible to obtain informed consent from data donors. They rely on a template

proposed by Sloan et al. (2020) listing the information that should be provided for consent to be informed. It includes: (1) why data is being collected; (2) what will be done with the data; (3) what data will be collected; (4) how data will be stored; and (5) what the risks of disclosure might be. Although such information is important and must be provided, we argue it is not enough for people to be *adequately* informed; that is, to have a clear understanding on *what* and *how* their data will be used (Neisse et al. 2016). The information provided is the mere formalization of a unilateral transaction of data, an element that remains opaque and abstract. Stating *what* data will be collected is widely different from supporting people in knowing and understanding data and its implications.

Our results underline how the content of personal data is not only opaque to donors but also to us, researchers. We were technically equipped to understand and analyze the received speech records and prepared to encounter contextual insights. However, we could not have anticipated the information unravelled through the process. Thus, we propose an iterative and incremental process of supporting participants in *knowing* (what) data that invites them to (re) evaluate and (re)assess their participation, preferences, and privacy boundaries. Hence, improving the informed consent process towards one that is ongoing and dynamic (Kaye et al. 2015). This process requires researchers and designers to reconsider their relationships with participants and adopt new procedures that harness the dynamic nature of the data, continuously changing through the actions and preferences of participants (Gómez Ortega et al. 2022).

5.2 Understanding as value gain

Previous research argues for ensuring that data donors derive value from engaging in data donation (Prainsack 2019a; Hummel et al. 2019; Bietz et al. 2019). They have proposed potential avenues for it; including (1) positive feelings derived from contributing to scientific research (Skatova et al. 2014; Skatova and Goulding 2019), and (2) future benefits from the research outputs (Prainsack 2019a; Skatova and Goulding 2019). In this paper, we hypothesized that a better understanding (i.e., increased (personal) data literacy) of the data yields a return of value for donors. Our results illustrate that most donors perceive the incremental knowledge regarding their (personal) data as valuable, even if uncomfortable or creepy. In addition, most donors appreciated the *empowerment* derived from acquiring new knowledge that can be applied to other contexts and gaining *ownership* of their data.

We argue that fostering a better understanding of the (donated) data on a personal (i.e., how it relates to and reflects their behavior) and infrastructural (i.e., how data is collected, stored, and regulated) level is a promising avenue for donors to gain value from engaging in data donation. It is a value-gain strategy that harnesses the abilities and strengths of researchers (e.g., shaping the data and translating it into something graspable). It can support (better) informed transactions of and collaborations through data. Additionally, it can trigger researchers and designers to better understand and engage with people's entanglements with their data. These, in turn, can invite us as researchers to reflect on our practice (e.g., *how do we do research?*) (Gould 2022) and prompt us to design digital and AI systems that invite different relationships with (personal) data. Hence, it is an opportunity for mutual benefit that is feasible and relevant for both parties.

6 Data donation: beyond data transactions

Current framings of data donation (e.g., Skatova et al. 2014; Razi et al. 2022; Ohme and Araujo 2022) are focused mainly on the moment of the voluntary transaction of data (i.e., when a person donates her data). As part of this transaction, it is important to ensure that data donors: (1) are *adequately informed* regarding their participation (Ohme and Araujo 2022; Bietz et al. 2019; Jones 2019; Strotbaum et al. 2019) and (2) *gain value* from engaging in data donation (Prainsack 2019a; Hummel et al. 2019; Bietz et al. 2019).

We propose a framework that conceives data donation as an *encompassing process* around the voluntary transaction of data. This process, *by design*, should dynamically and iteratively support and invite donors to:

1. Access meaningful information about how their data is used and handled.
2. Explore and understand their data on a personal and infrastructural level.
3. (Re)evaluate and (re)assess their boundaries and participation.

In Sect. 5.1, we argued for approaching informed consent as ongoing and dynamic; this is intrinsic and fundamental to a data donation process. Approaching data donation as an encompassing process de-emphasizes the transaction of data as the primary instance of informed consent. Yet, it still is the entry point into the process. We recommend concrete actions around this instance. *Before* the transaction of data, researchers and designers should inform donors about our goals and activities and enable them to explore and familiarize themselves with the data; grasp its content and characteristics. In this paper, we encouraged exploration *before* the transaction of data through an interactive graph where donors could engage with their speech records over time. *After* the transaction of data, and *throughout* the process, researchers and designers should support donors'

incremental understanding of the data. We supported the incremental understanding of the data by facilitating interpretation and articulation during semi-structured interviews prompted by the *personal data canvas*. Although these are not the only ways, they illustrate how these activities could manifest in practice.

In addition, we recommend researchers and designers to remain available and accountable to data donors. Accountability is important as data donation is a "leap of faith", as described in Sect. 4.5. That is, once researchers and designers gain access to the donated data, nothing (beyond ethics) prevents them from using it in a different way than agreed. Therefore, accountability could provide reassurance and increase donors' trust in the process.

6.1 Relationships between receivers, donors, and data

Our framework involves data as central entity, along with data donors and receivers as stakeholders. Our contribution lies in the relationship between these three elements.

1. *Data receivers* (i.e., researchers and designers) initiate data donation by inviting people to participate in their research. The data donation process requires them to intentionally engage with the personal and dynamic nature of the data. We recommend that, as part of their study design or design process, they contemplate: (1) facilitating data exploration and interpretation; (2) supporting donors in (re)evaluating their participation and (re)defining their boundaries (e.g., curating their data); (3) creating opportunities for donors to gain value; and (4) nurturing communication and accountability. In turn, these activities could contribute to their understanding of people's relations with their data and trigger reflection in their research and practice around digital and AI systems.
2. *Data donors* enable data donation by transferring their data. The data donation process invites them to actively contribute to research. We recommend that, as part of the process, they have the opportunity to explore their (personal) data and gain incremental and situated knowledge about its content. Through these activities donors gain a return of value through a better understanding of their data (i.e., increased (personal) data literacy). They gain ownership of their data, from being about them to theirs (i.e., they own an actual copy). Furthermore, they are empowered to make better-informed decisions throughout the research and beyond. Moreover, as described by previous research (e.g., Gómez Ortega et al. (2022); Ohme and Araujo (2022)), data donation should support their agency and autonomy by enabling them to set granular boundaries and define the terms of their participation.
3. *Data* evolves through data donation. Understanding data donation as an encompassing process implies understanding data as dynamic rather than static (always accessible and reusable). The data donation process is centered around the voluntary transaction of personal data and is further shaped by its exploration and understanding. Through data donation, personal data goes from an abstract entity to a concrete one that is situated and deeply entangled with people's behaviors and intimately relates to and reflects them. We encourage data donors and receivers to harness the dynamic nature of data, as it enables meaningful collaborations shaped by both parties and from which both parties benefit.

6.2 Critical role of data literacy in data donation

Data literacy is a widely used concept often associated with specific skills and abilities (e.g., combining data, visualizing data). We provide recommendations to foster data donors' (personal) data literacy. Concretely, we focus on the information that should be communicated to support informed data donating (and broadly data sharing) decisions.

1. *How is data collected?* Provide information on data collection (and storage) and its relationship to people's behavior and interaction with digital products and services. For example, a speech record is generated and stored every time a person interacts with her Google Assistant.
2. *How can I access my data?* Provide information regarding specific data collection practices and policies that support people in navigating the process of gaining *ownership* of their data. For example, speech records collected by Google Assistant are available upon request via Google Takeout.
3. *What exactly is (in) my data?* Enable playful ways for people to understand the content and characteristics of the data, what it *looks*, *sounds*, and *feels* like. For example, illustrate how often a data point is generated, what information in contains, and how many are on the entire dataset.
4. *What makes data about me?* Illustrate how data relates (and reflects) to people, their behavior, and experiences. For example, by shaping data in a way that underlines behavioral patterns and facilitates interpretation.
5. *What data I'm donating?* Support awareness of the data (and personal information) being shared and its potential implications. Enable and facilitate setting boundaries and identifying potentially sensitive elements. For example, allow for granular data-sharing decisions throughout the process.

6.3 Limitations

Although data donation offers new ways of engaging with and accessing (personal) behavioral data; our research underlines certain limitations and important considerations. First, data donation is limited by who is able and willing to donate their data. A group of donors might likely be different from one recruited through different means or responding to different incentives (e.g., money, rewards). Similarly, our study was limited by the people who could donate their data, the framing and channels we used to disseminate our call to donate, and the types of data we requested. Second, data donation is highly dependent on local regulations, and the degree to which they enable individuals to obtain a copy of their data as well as how well they are enforced. Third, data donation is shaped by personal interaction patterns (i.e., how people interact with a product and service, which shapes the frequency, amount, and types of data) and configuration settings (i.e., how people configure a product or service, which shapes the availability of data). Fourth, due to data being *relational* (Prainsack 2019b), data donation could entail the transaction of data from people other than the person actively participating in (and consenting to) the research. We should be mindful of the relationality of the data and find ways for these people to actively participate in the research or be excluded from the (donated) data. Finally, we propose a data donation framework that dynamically supports informed consent, resulting in the generation of highly dynamic datasets. Future research should investigate how the dynamic nature of the data fits within existing paradigms and practices (e.g., FAIR principles, open data).

7 Conclusion

In this paper, we adopted a Research through Design (RtD) approach to investigate how to foster (personal) data literacy and increase donors' understanding of their data. We defined, designed, and developed a data donation journey mediated by a data donation platform. Throughout the journey, we invited 22 Google Assistant users (i.e., donors) to donate their speech records and reflect on their understanding of the data and their data donation experience via short questionnaires and semi-structured interviews. Our findings indicate that supporting an incremental understanding of the data enables donors to be *adequately* informed. Additionally, the knowledge and empowerment derived from the increased understanding of the data and how it relates to and reflects a person's behavior are perceived as valuable. We offer our recommendations in the form of a framework that conceives data donation as an integral process around voluntary transactions of data where donors are supported in exploring and understanding their data and encouraged to (re)evaluate and (re)assess their participation. Through this process, we propose to situate informed consent in a broader set of activities before, during, and after the voluntary transactions of data.

Data availability The datasets generated and analyzed during the current study are not publicly available as research participants did not provide their informed consent for this purpose.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Alizadeh F, Jakobi T, Boldt J et al (2019) GDPR-reality check on the right to access data. In: Proceedings of Mensch und Computer 2019. ACM, New York, pp 811–814, https://doi.org/10.1145/3340764.3344913

Andreotta AJ, Kirkham N, Rizzi M (2021) AI, big data, and the future of consent. AI & Soc. https://doi.org/10.1007/s00146-021-01262-5

Bentley F, Luvogt C, Silverman M et al (2018) Understanding the long-term use of smart speaker assistants. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2(3):1–24. https://doi.org/10.1145/3264901

Bietz M, Patrick K, Bloss C (2019) Data Donation as a Model for Citizen Science Health Research. Citizen Science: Theory and Practice 4(1):1–11. https://doi.org/10.5334/cstp.178

Bogers S, Frens J, van Kollenburg J, et al (2016) Connected baby bottle. In: Proceedings of the 2016 ACM Conference on Designing Interactive Systems. ACM, New York, pp 301–311, https://doi.org/10.1145/2901790.2901855,

Bourgeois J, van der Linden J, Kortuem G et al (2014) Conversations with my washing machine. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, New York, pp 459–470, https://doi.org/10.1145/2632048.2632106

Bowyer A, Holt J, Go Jefferies J et al (2022) Human-GDPR interaction: practical experiences of accessing personal data. In: CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–19, https://doi.org/10.1145/3491102.3501947

Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3(2):77–101. https://doi.org/10.1191/1478088706qp063oa

Braun V, Clarke V (2013) Successful qualitative research, a practical guide for beginners. SAGE Publications Ltd, London

Breuer J, Kmetty Z, Haim M et al (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 pp 1–20. https://doi.org/10.1080/1369118X.2022.2097015

Brown B, Weilenmann A, McMillan D et al (2016) Five provocations for ethical HCI research. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 852–863, https://doi.org/10.1145/2858036.2858313

Chalhoub G, Kraemer MJ, Nthala N et al (2021) "It did not give me an option to decline": a longitudinal analysis of the user experience of security and privacy in smart home products. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–16, https://doi.org/10.1145/3411764.3445691

Choe EK, Lee B, Andersen TO et al (2018) Harnessing the power of patient-generated data. IEEE Perv Comput 17(2):50–56. https://doi.org/10.1109/MPRV.2018.022511243

Clegg T, Greene DM, Beard N, et al (2020) Data everyday: data literacy practices in a division i college sports context. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–13, https://doi.org/10.1145/3313831.3376153

Cooper D, Ubben T, Knoll C et al (2022) Open-source web portal for managing self-reported data and real-world data donation in diabetes research: platform feasibility study. JMIR Diabetes 7(1):e33213. https://doi.org/10.2196/33213

Diethei D, Niess J (2021) Sharing heartbeats: motivations of citizen scientists in times of crises. In: Conference on Human Factors in Computing Systems—Proceedings 2020(April):15. https://doi.org/10.1145/3411764.3445665

GDPR (2018) General data protection regulation. https://gdpr.eu/

Giaccardi E, Stappers PJ (2017) Research through design. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/research-through-design

Gómez Ortega A, Bourgeois J, Kortuem G (2022) Reconstructing intimate contexts through data donation: a case study in menstrual tracking technologies. In: Nordic Human-Computer Interaction Conference. ACM, New York, pp 1–12, https://doi.org/10.1145/3546155.3546646

Gorkovenko K, Burnett DJ, Thorp J et al (2019) Supporting real-time contextual inquiry through sensor data supporting real-time contextual inquiry through sensor data. In: Ethnographic Praxis in Industry Conference Proceedings. Edinburgh, UK, pp 1–29

Gould SJJ (2022) Consumption experiences in the research process. In: CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–17, https://doi.org/10.1145/3491102.3502001

Gray J, Gerlitz C, Bounegru L (2018) Data infrastructure literacy. Big Data Soc. https://doi.org/10.1177/2053951718786316

Hummel P, Braun M, Dabrock P (2019) Data donations as exercises of sovereignty. In: Philosophical Studies Series, vol 137. Springer International Publishing, Cham, p 23–54, https://doi.org/10.1007/978-3-030-04363-6_3

Jakobi T, Stevens G, Castelli N et al (2018) Evolving needs in IoT control and accountability. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2(4):1–28. https://doi.org/10.1145/3287049

Jones KH (2019) Incongruities and dilemmas in data donation: juggling our 1s and 0s. Philos Stud Ser 137:75–93. https://doi.org/10.1007/978-3-030-04363-6_5

Kaye J, Whitley EA, Lund D et al (2015) Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 23(2):141–146. https://doi.org/10.1038/ejhg.2014.71

Krutzinna J, Taddeo M, Floridi L (2019) An ethical code for posthumous medical data donation. Philos Stud Ser 137:181–195. https://doi.org/10.1007/978-3-030-04363-6_12

Kurze A, Bischof A, Totzauer S, et al (2020) Guess the data: data work to understand how people make sense of and use simple sensor data from homes. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–12, https://doi.org/10.1145/3313831.3376273

Kwon H, Fischer JE, Flintham M et al (2018) The connected shower. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2(4):1–22. https://doi.org/10.1145/3287054

Lau J, Zimmerman B, Schaub F (2018) Alexa, are you listening? In: Proceedings of the ACM on Human-Computer Interaction 2(CSCW):1–31. https://doi.org/10.1145/3274371

Liao M, Sundar SS (2021) How should AI systems talk to users when collecting their personal information? Effects of role framing and self-referencing on human-AI interaction. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–14, https://doi.org/10.1145/3411764.3445415

Low DM, Bentley KH, Ghosh SS (2020) Automated assessment of psychiatric disorders using speech: a systematic review. Laryngosc Investig Otolaryngol 5(1):96–116. https://doi.org/10.1002/lio2.354

Malkin N, Deatrick J, Tong A et al (2019) Privacy attitudes of smart speaker users. Proc Priv Enhanc Technol 4:250–271. https://doi.org/10.2478/popets-2019-0068

Martelaro N, Lakdawala T, Chen J, et al (2021) Leveraging the twitch platform and gamification to generate home audio datasets. In: Designing Interactive Systems Conference 2021. ACM, New York, pp 1765–1782, https://doi.org/10.1145/3461778.3462097

Maus B, Salvi D, Olsson CM (2020) Enhancing citizens trust in technologies for data donation in clinical research: validation of a design prototype. In: 10th International Conference on the Internet of Things Companion. ACM, New York, pp 1–8, https://doi.org/10.1145/3423423.3423430

Neisse R, Baldini G, Steri G, et al (2016) Informed consent in Internet of Things: the case study of cooperative intelligent transport systems. In: 2016 23rd International Conference on Telecommunications (ICT). IEEE, pp 1–5, https://doi.org/10.1109/ICT.2016.7500480

O'Connor Y, Rowan W, Lynch L et al (2017) Privacy by design: informed consent and internet of things for smart health. Procedia Comput Sci 113:653–658

Ohme J, Araujo T (2022) Digital data donations: a quest for best practices. Patterns 3(4):100467. https://doi.org/10.1016/j.patter.2022.100467

Pins D, Jakobi T, Boden A, et al (2021) Alexa, we need to talk: a data literacy approach on voice assistants. In: Designing Interactive Systems Conference 2021. ACM, New York, pp 495–507, https://doi.org/10.1145/3461778.3462001

Prainsack B (2019) Data donation: how to resist the iLeviathan. Philos Stud Ser 137:9–22. https://doi.org/10.1007/978-3-030-04363-6_2

Prainsack B (2019) Logged out: ownership, exclusion and public value in the digital data and information commons. Big Data Soc 6(1):205395171982977. https://doi.org/10.1177/2053951719829773

Pu X, Kross S, Hofman JM, et al (2021) Datamations: animated explanations of data analysis pipelines. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–14, https://doi.org/10.1145/3411764.3445063

Razi A, Alsoubai A, Kim S, et al (2022) Instagram data donation: a case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. In: Conference on Human Factors in Computing Systems - Proceedings, pp 1–18, https://doi.org/10.1145/3491101.3503569

Robert Koch Institut (2020) Corona-Datenspende. https://corona-datenspende.de/science/en/

Rudnicka A, Cox AL, Gould SJJ (2019) Why do you need this? In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–11, https://doi.org/10.1145/3290605.3300622

Sciuto A, Saini A, Forlizzi J, et al (2018) "Hey Alexa, what's up?". In: Proceedings of the 2018 Designing Interactive Systems Conference. ACM, New York, pp 857–868, https://doi.org/10.1145/3196709.3196772

Skatova A, Goulding J (2019) Psychology of personal data donation. PLoS One 14(11):e0224240. https://doi.org/10.1371/journal.pone.0224240

Skatova A, Ng E, Goulding J (2014) Data donation: sharing personal data for public good? In: Digital Economy All Hands Meeting, December, pp 1–3, https://doi.org/10.13140/2.1.2567.8405

Sloan L, Jessop C, Al Baghal T et al (2020) Linking survey and twitter data: informed consent, disclosure, security, and archiving. J Empir Res Hum Res Ethics 15(1–2):63–76. https://doi.org/10.1177/1556264619853447

Strotbaum V, Pobiruchin M, Schreiweis B et al (2019) Your data is gold—data donation for better healthcare? Inform Technol 61(5–6):219–229. https://doi.org/10.1515/itit-2019-0024

Tolmie P, Crabtree A, Rodden T, et al (2016) "This has to be the cats". In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. ACM, New York, CSCW '16, pp 491–502, https://doi.org/10.1145/2818048.2819992

US Chambers of Commerce Foundation (2014) The Future of Data-Driven Innovation. Tech. rep., U.S. Chambers of Commerce Foundation, https://www.uschamberfoundation.org/sites/default/files/The Future of Data-Driven Innovation.pdf

Van den Berg B, Van der Hof S (2012) What happens to my data? A novel approach to informing users of data processing practices. First Monday 17(7):1–15. https://doi.org/10.5210/fm.v17i7.4010

Wolff A, Gooch D, Cavero Montaner J, et al (2017) Creating an understanding of data literacy for a data-driven society. J Commun Inform 12(3):(In press). www.ci-journal.net/index.php/ciej/article/view/1286

Zimmerman J, Forlizzi J, Evenson S (2007) Research through design as a method for interaction design research in HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, pp 493–502, https://doi.org/10.1145/1240624.1240704