Data Subjects' Reactions to Exercising Their Right of Access

Arthur Borem, Elleen Pan, Olufunmilola Obielodan, Aurelie Roubinowitz, and Luca Dovichi, *University of Chicago;* Michelle L. Mazurek, *University of Maryland;* Blase Ur, *University of Chicago*

https://www.usenix.org/conference/usenixsecurity24/presentation/borem

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

Data Subjects' Reactions to Exercising Their Right of Access

Arthur Borem, Elleen Pan, Olufunmilola Obielodan, Aurelie Roubinowitz, Luca Dovichi, Michelle L. Mazurek†, Blase Ur
University of Chicago and †University of Maryland

Abstract

Recent privacy laws have strengthened data subjects' right to access personal data collected by companies. Prior work has found that data exports companies provide consumers in response to Data Subject Access Requests (DSARs) can be overwhelming and hard to understand. To identify directions for improving the user experience of data exports, we conducted an online study in which 33 participants explored their own data from Amazon, Facebook, Google, Spotify, or Uber. Participants articulated questions they hoped to answer using the exports. They also annotated parts of the data they found confusing, creepy, interesting, or surprising. While participants hoped to learn either about their own usage of the platform or how the company collects and uses their personal data, these questions were often left unanswered. Participants' annotations documented their excitement at finding data records that triggered nostalgia, but also shock about the privacy implications of other data they saw. Having examined their data, many participants hoped to request the company erase some, but not all, of the data. We discuss opportunities for future transparency-enhancing tools and enhanced laws.

1 Introduction

Over the past decade, legislators in many countries have strengthened consumer data protection and privacy rights. Two prominent privacy laws are the E.U.'s General Data Protection Regulation (**GDPR**) [21] and the California Privacy Rights Act (**CPRA**) [11]. Both aim to strengthen consumers' control over data collection and processing, influencing subsequent legislation in Japan [15], Utah [60], Connecticut [54], and Colorado [26]. The GDPR in particular guarantees four key rights: knowledge, deletion, opt-out, and non-discrimination. We focus on the right to knowledge, and specifically whether online platforms are respecting this right.

This right to knowledge, or access, is a critical component of U.S. and international privacy regulation and has been for decades, even before the GDPR and the CPRA. The U.S. Privacy Act of 1974 mandates that individuals can "gain access to [their] record or to any information pertaining to [them]" for government-held data [58]. The right to access is one of nine Fair Information Practice Principles (FIPPs) [23] and eight OECD Privacy Principles [27]. This decades-old right is a key mechanism of privacy transparency. By examining their data, individuals can evaluate an organization's data practices and take action, such as by deleting or modifying data the organization holds about them.

A primary vehicle for guaranteeing the right to knowledge is the Data Subject Access Request (**DSAR**), a communication "addressed to the organization that gives individuals a right to access information about personal data the organization is processing about them" [2, 16]. Once a DSAR is made, consumers receive a **data export**, consisting of one or more files (potentially of many types) containing their personal data. Article 12 of the GDPR specifies that these data exports must be sent "without undue delay" (at most "within one month of receipt of the request") and be "concise, transparent, intelligible and easily accessible" [21]. Further, Article 20 requires that these exports be portable, meaning they must be in a "structured, commonly used and machine-readable format."

Previous work has shown it is challenging to structure a data export in a way that is machine-readable (Article 12), yet still intelligible to consumers (Article 20). Veys et al. conducted focus groups of consumers and showed that companies are not overcoming this challenge [62]. Their participants found data exports unusable or otherwise not useful in their current state. This state of affairs is unfortunate. Prior studies have demonstrated that consumers are interested in, and can gain value from, data exports [4, 62, 64]. Furthermore, tools like Spotify Wrapped, which visualize personal data (not necessarily from data exports), have become popular [46, 53, 55]. Unfortunately, these tools typically visualize only a handful of metrics. For instance, Spotify Wrapped displays a user's music-listening habits [53], but a data export from Spotify contains far more information. A recent research prototype, TransparencyVis [49], aims to enable general-purpose interactions with an entire data export, but does not appear to have been designed based on users' goals for data exports.

Thus, we argue now is a good time to (1) determine to what degree exports are adhering to this legislation, (2) determine the degree to which current data exports are meeting consumers' needs, and (3) pinpoint ways of refining future legislation by identifying what consumers are most interested in learning from their exports. To this end, we conducted a user study in which 33 users of Amazon, Facebook, Google, Spotify, and Uber made DSARs, explored and annotated their data exports using a web app we built, shared pseudonymized versions of their exports, and reflected on their experiences.

In this paper, we answer the following research questions:

RQ1: How complex or concise are data exports? The GDPR states data exports should be "concise" and (along with the CPRA in identical language) presented in a "structured, commonly used and machine-readable format." Concise exports are more accessible for consumers, making it so that they can enact their privacy choices without the burden of extensive searching. To answer this question, we analyzed the 801 data export files 33 participants shared. We found that the complexity of data exports varied both across and within platforms. Amazon data exports typically contained a dozen directories; Spotify data exports contained only one.

Any consumer or data subject should be able to use data exports as a window into a company's privacy practices. The GDPR codifies this by requiring that the information in data exports be in an "intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child." The CPRA similarly states that data exports should be "easily understandable to the average consumer." Therefore, we ask:

RQ2: What kinds of questions did participants want to answer with their data exports, and which of these were they eventually able to answer? To investigate, we analyzed questions our 33 participants submitted before exploring their data, alongside their reflections on those questions. Participants were most curious to learn about their usage patterns (e.g., hours spent daily), platforms' data-collection practices, and how the platforms use personal data. Because of the heavy use of jargon, large exports, and specialized file formats, participants were often left with incomplete or missing answers.

RQ3: How effective are data exports as tools for providing information to consumers? Since data exports are inaccessible to most consumers due to their size, file format, and complexity [62], we built a React-based web application with which participants explored their exports and annotated bits of information they found confusing, creepy, interesting, or surprising. These annotations showed that participants reacted positively to finding records that triggered memories. They also liked finding record types for which they searched (e.g., login times). Notably, they looked for a wide variety of records. For example, while some wanted to map out every Uber ride they had ever taken, others wanted to discover new music based on past Spotify listening behavior. In contrast, a number of participants were shocked by the records they saw, feeling their privacy had been violated and their agency undermined. Some participants were especially surprised when they discovered data they thought they had previously deleted.

In addition to the right of access, the GDPR and CPRA give consumers the right to modify incorrect data, erase data, and restrict processing of data for specific purposes (e.g., prevent the sale of personal data to third parties). These rights are how consumers enact their data privacy preferences and are most useful once consumers have the knowledge granted by data access rights. To evaluate their relevance, we ask:

RQ4: Are the actions consumers are most interested in taking after exploring their data in line with those guaranteed by privacy laws? In a post-task survey, participants wanted specific data records erased if they could not justify why the platform would collect that record. In their annotations, participants similarly wondered how they could prevent the platform from collecting analogous data in the future.

Our findings suggest the need for redesigning platforms for exploring data exports to better meet consumer needs. These tools should include a comprehensive list of all data record types in the export. More often than not, basic syntheses and visualizations of each record type will be sufficient. To more closely abide by Article 12 of the GDPR, companies should experiment with different data formats for their data exports and define jargon any time it is used. Finally, data exports could promote user agency and trust by embedding actions consumers can take (e.g., modifying a specific setting or requesting a record be erased) within the exploration process.

2 Background and related work

Data subject rights and security and privacy. Since the GDPR went into effect, the security and privacy community has been studying the impacts of data subject rights. Many studies have found shortcomings of the GDPR, mostly due to imprecise language, and have developed solutions. Some were focused on challenges with implementation, such as Cohen et al., who found that absolute deletion has complex implications for machine learning models that have already been trained on that data [13]. They instead proposed leveraging differential privacy mechanisms to guarantee near indistinguishability between models where the data subject is present and those where the data subject is absent [13]. Ferreira et al. developed a system for guaranteeing data protection requirements that can easily be added to existing web applications [24]. Our study takes the complementary approach of exploring current implementations of the right to access and identifying the core features an effective implementation should have.

Others have studied how end users interface with privacy regulation. Habib et al. found that data deletion options and opt-outs for email communications and targeted advertising are not easy to use [29]. Utz et al. argued that the methods for obtaining consent for cookies should be made more explicit in existing regulation since small implementation de-

cisions (e.g., the positions of cookie banners) have a large impact on user actions [61]. Farke et al. showed that privacy dashboards—interfaces that allow users to review and control data collection that partially parallel the right of access [45]—strengthen trust between users and platforms and are therefore effective tools for communicating security and privacy data practices [22]. Our study similarly asks if the right to access as it currently exists is sufficient to meet consumer needs.

Some studies have focused on identifying adherence to—or violations of—privacy regulations. Degeling et al. analyzed privacy changes on popular websites after the GDPR's introduction, finding a 16% increase in cookie consent notices on European websites [19]. Bertram et al. analyzed URLs users requested to be removed from Google Search for five years following the passage of a "right to be forgotten," demonstrating that consumers were enacting their right and Google was complying [5]. Nguyen et al. found that nearly 30% of Android apps sent personal user data to data controllers without the user's explicit prior consent [39]. In our study, we search for similar instances where platforms could be more compliant with consumers' right of access.

Data exports. Prior work has examined the DSAR process and the resultant data exports. Initial work highlighted vulnerabilities in the DSAR authentication process that can be abused to steal data from others [7, 36]. Researchers have also shown that consumers experience long wait times [18, 32] and encounter manipulation tactics [33], poorly organized exports, and privacy risks during authentication [10, 32, 52, 57, 59].

Other work has leveraged data exports for purposes beyond data subjects' rights [30]. For instance, researchers have used data exports (as "data donations") to audit social media platforms like Instagram [43], TikTok [65], and Twitter [64].

Academics and industry practitioners have also explored ways of improving DSARs. Leschke et al. designed an automated approach to making DSARs, minimizing the burden on consumers [34]. Others have attempted to streamline a company's process of responding to DSARs by either providing tooling for data workers [47, 56], attempting to automate the process [31, 38, 63], or studying inefficiencies in GDPR compliance as a whole [35, 48, 50, 51].

Making sense of personal data. Information about a company's privacy policies and privacy-enhancing features can be hard to find and is often written in complex, inaccessible ways [40, 42, 44]. Previous work has shown that not only do consumers have an interest in enhancing their privacy [14, 20, 37], but tools and interfaces that increase transparency into company data practices lead to increased interest and concern with privacy. For example, Arias-Cabarcos et al. showed that Meta's "Off-Facebook Activity" Dashboard was effective at informing consumers of previously unknown flows of their data [4]. Datta et al. built a similar interface for Google's Ad Settings page [17]. Angulo et al. built a tool for consumers to visualize data disclosures, which they similarly found to be clearly informative [3]. Other work has demonstrated the effectiveness of privacy nudges [1], data disclosure dashboards [3, 22, 25], mobile privacy dashboards [41], and data flow dashboards [6]. Most relevant to this study is Schufrin et al.'s TransparencyVis prototype, developed for visualizing data exports [49]. Our investigation into users' goals complements their tool and suggests future directions.

3 Method

To understand users' goals and reactions to exploring their data exports, we conducted a two-part online user study. The first part was a screening survey where we asked participants to make DSARs for one or more of the five companies studied (see Section 3.1). The main study consisted of participants exploring and annotating their own data export. The screening survey needed to be separate from the main study for two reasons. First, platforms do not respond to DSARs immediately. Second, we wanted to filter for participants who had a significant history of usage with the companies chosen.

In the main study, participants uploaded their data to an interactive web platform we built. We decided to build such a platform because, as previous research has shown [62], most consumers are not familiar with reading or opening the files in formats many platforms use in their data exports (e.g., JSON). Additionally, since we wanted to record detailed and accurate reactions, we wanted to collect these *while* participants were exploring their data. Therefore, our platform included rich annotation features (see Section 3.3).

3.1 Platform selection and DSAR instructions

We recruited users of five platforms for this study: Amazon, Facebook, Google, Spotify, and Uber. We selected these platforms because they are widely used and because they have been a focus of previous work on data exports [49, 62]. Our original study design also included Twitter (currently rebranded as X), but both the research team and pilot testers experienced substantial or indefinite delays in receiving their data exports, so we needed to exclude it.

Because of technical constraints described in Section 3.3, our annotation platform could not always accept participants' entire data exports, which for some companies can include tens of gigabytes of data for active users. Fortunately, with the exception of Uber, the platforms we chose allow consumers to request a specific subset of their personal data. When the option existed, we instructed participants to request their data in JSON or CSV format, as opposed to HTML format. The research team and pilot testers submitted several requests to the five companies throughout the research process to determine the approximate size and types of files contained within these exports. We then used this information to identify categories to exclude from Amazon, Facebook, Google, and Spotify's

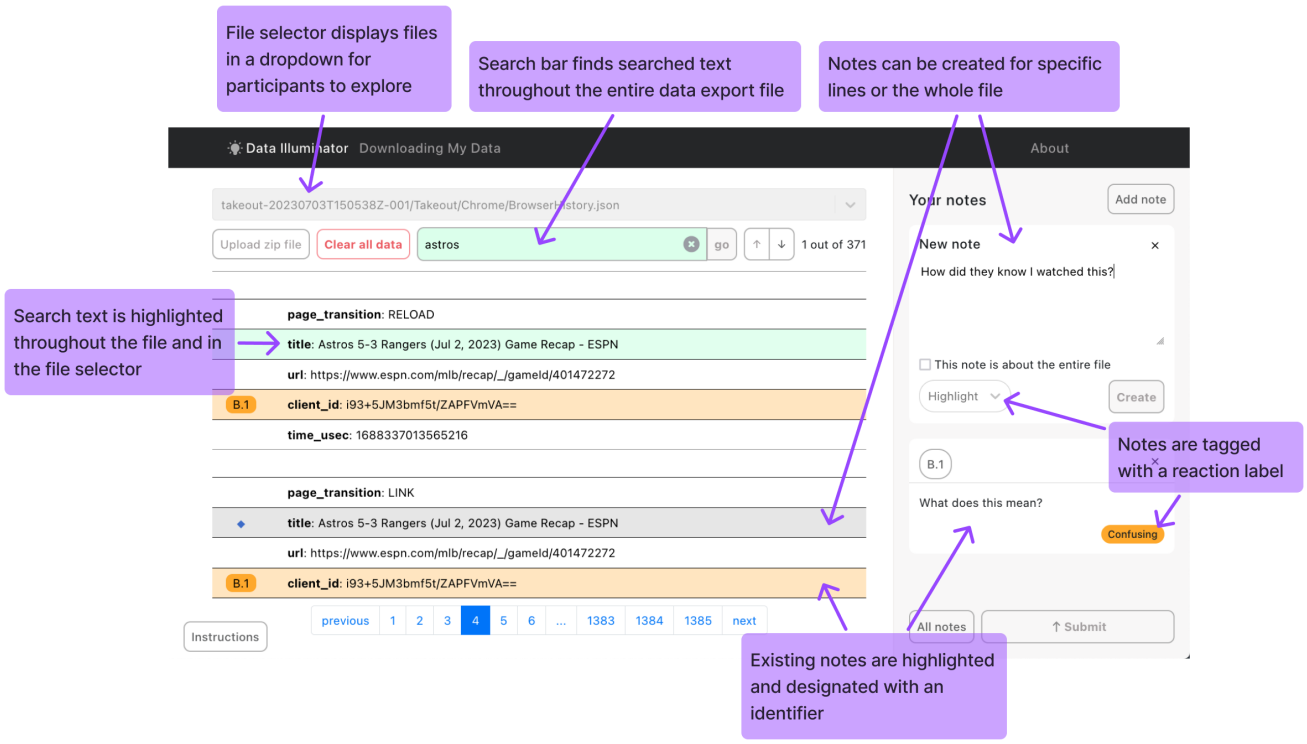


Figure 1: Diagram showing the process of interacting with the web-based tool we built for exploring and annotating data exports.

DSARs. These excluded categories included large, sensitive data like media (photos and videos) and documents stored in Google Drive, as well as the extended streaming history for Spotify. Despite these precautions, some participants' data exports were still too large for our platform. In these cases, we worked with them to identify folders and files that were requested by accident or unusually large (many gigabytes).

3.2 Recruitment and screening survey

We recruited participants on the Prolific crowdwork platform. The screening task was open to any English speaker in the world over the age of 18. Recruiting from an international pool meant our participants completed the study under a variety of privacy regulatory regimes. We did so to ensure our eventual recommendations to tool developers and policymakers would benefit a wide range of consumers. Recruiting from multiple countries and languages provided visibility into the structure of data exports from around the world. This strategy also helped with recruiting more participants for our complex and lengthy study. We directed participants to a Qualtrics survey that asked about their usage habits of the selected companies. Any participant who had been using at least one of these platforms for at least two years was given instructions on submitting a DSAR. We then asked participants to share their experiences requesting their data. All participants were paid $2 USD for completing the screening survey regardless of whether they qualified for the main study.

3.3 Main study task

In the main study, we first asked participants in a Qualtrics survey to confirm they received their data export and to answer questions about their previous experiences making DSARs and exploring data exports. We then asked them to brainstorm five questions they hoped could be answered using their data export. At this point, we asked participants to upload their data export to our web app and begin the annotation task.

We built our data annotation platform using Node.js and React, hosting it on our research group's servers. Figure 1 shows the main interface and features. After they uploaded their data exports, we asked participants to create 10 annotations on the platform (or five if their data export had 20 or fewer files). An **annotation** consists of three elements: zero or more lines within a file, an explanation of why those lines or that file stood out, and one of four tags—**confusing**, **creepy**, **interesting**, or **surprising**—proposed by Veys et al. [62].

Once participants created the minimum number of annotations, they had the option of creating more or beginning the submission process. In the first step in that process, our platform showed the participants their five original questions and asked them to reflect on whether or not they had been answered. Then, the data annotation platform prompted participants to identify three additional questions they had about themselves, their data, or the company that provided the data export. We paid participants $15 USD for completing the main study, which took about 35 minutes. We allowed each

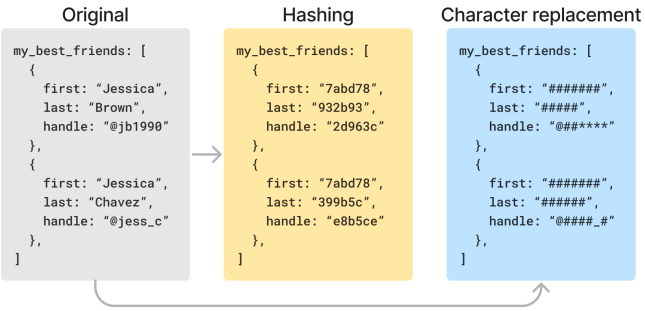


Figure 2: We created two anonymized versions of the data export using hashing and character replacement, respectively.

participant to complete the annotation task for more than one company, compensating them separately for each.

3.4 Ensuring participant privacy

In addition to receiving approval from the University of Chicago Institutional Review Board (IRB), we took a series of precautions to protect participants' privacy and anonymity. The design of our study required participants to upload data that, by its nature, contained identifying information. A data export might include emails, full names, addresses, dates of birth, and more. It also likely contains information that participants would consider sensitive and may not want to share, such as their location history or browsing history.

We used two primary strategies to protect participants' privacy. First, we performed all processing on the client side. Second, we pseudonymized the data exports before uploading them to our server. As mentioned earlier, the research team completed several DSARs for the five companies and found that the keys in JSON data never contained user data—instead they were used to name types of records. Therefore, we only pseudonymized the corresponding values in the exports.

We performed **hashing** and **character replacement** on the values to keep important structural data while maintaining participant privacy (see Figure 2). When hashing, we first split values by white space, leaving us with a list of tokens. We then salted each token on the client side with a randomly generated 32-byte string unique to each participant, which was never shared with the researchers and would be computationally infeasible to brute force. Our web app uploaded to our access-controlled server, via HTTPS, a version of the data export in which the keys were retained in plain text, but all values had been salted and then hashed. This hashed version allowed us to identify repeated values since a given term with a fixed salt always hashes to the same value. Our web app also created a separate version of the data export in which values underwent character replacement. We replaced all letters in the values with pound characters (#) and all digits with asterisks (*). Character replacement preserved the file size and the length of values, enabling additional analyses. Appendix A details how we communicated these strategies to prospective participants.

One participant (P20) reached out to us on Prolific with concerns about their data privacy and security. They were worried about exposing financial information to us, our ability to access their data, and whether they might accidentally delete data by making a DSAR. We responded with the same description of the anonymization techniques we put in the tool. The participant thanked us, said they were satisfied with our precautions, and completed the study.

3.5 Qualitative data analysis

We performed qualitative analysis on participants' free-text responses to identify themes. Two team members collected responses to the data annotation task and the survey from about 10% of participants and compared themes to develop a preliminary codebook. While this sufficed for identifying primary codes, all four coders subsequently iterated on secondary codes. We then split responses into four sections. One team member served as first coder on all responses, while four separate team members served respectively as second coder. Each pair met and resolved all disagreements. Finally, the first coder performed affinity diagramming on the codes to identify connections between themes. We correct spelling and capitalization mistakes when reporting participant quotes.

3.6 Limitations

Our study considered only a subset of the files and data that the selected companies have on users. This limitation was due to the technical challenges brought on by relying entirely on client-side computation, a trade-off we believe was important to ensure participant privacy and promote trust. Additionally, we only built capacity for parsing JSON files, CSV files, and (when the file structure was predictable) selected HTML files. All other files were discarded. While this does limit our analysis of the size and structure of participants' data exports, most participants did not click through every single file.

Among companies, Spotify and Amazon had the smallest number of participants in our study. We suspect this limitation was caused by these platforms taking a week or two, the longest among the chosen platforms, to respond to DSARs. Participants would sometimes forget they signed up for the study, not notice they received a data export email before the download link expired, or simply turn down the Prolific task.

While we anticipate that our hashing and redaction techniques should prevent most potential privacy leaks, there is no guarantee that identifying information cannot be found in the names of files or in JSON keys even though we never observed this to happen in our own data exports. For this reason, we made ourselves as available as possible to our participants via email and on the Prolific platform. Most participants did not reach out with privacy concerns, but we promptly responded to the one who did (P20); they chose to complete the study.

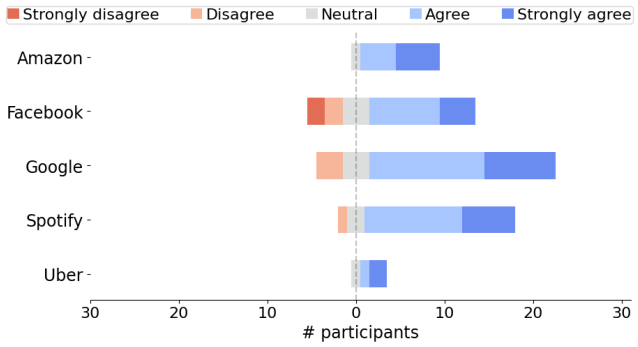


Figure 3: Participants' responses to the following statement: "It was easy to download my data from the platform."

Additionally, as Table 1 shows, our sample is not representative of global consumers and skewed young, white, and male relative to the broader population. Participants may know more about data practices, which could impact reactions to their personal data. In Section 3.3 we described how, in the screening survey, we asked participants to articulate five questions they hoped could be answered during the study using their data export. Some participants might not have had any questions about their data and thus generated questions for the sake of the task. As such, the questions some participants produced may reflect a set of questions they might abstractly hope to have answered somehow, rather than what they realistically expected to be answered by their data exports. Note, however, that the questions generated by the six participants who viewed the data requested for this study before the annotation task were not atypical of those of other participants: they were about general data usage, searching for specific records (e.g., location), and company security practices.

4 Results

We first describe our participants and their DSAR experiences (Section 4.1) before characterizing their data exports (Section 4.2). We then describe the questions participants posed about their data exports (Section 4.3) and key themes from participants' annotations (Section 4.4). Finally, we describe participants' intended future actions (Section 4.5).

4.1 Participants and their DSAR experiences

In total, 33 participants completed the study. Table 1 summarizes their demographics. Participants skewed young (24 younger than 35), male (21), and white (21). The most represented country was South Africa (10), and the most represented continent was Europe (15). Only 14 participants hold a bachelor's degree or higher, and the majority work outside the technology field. Recall that we permitted participants to complete the data-annotation task for multiple companies; five annotated data exports for two different companies.

Table 1: Participants' demographics and prior experiences.

Category	n	%
Gender		
Male	21	63
Female	9	27
Non-Binary	3	9
Technical work experience		
No	18	55
Yes	15	45
Country		
South Africa	10	30
Poland	5	15
Italy	3	9
USA	2	6
Portugal	2	6
Greece	1	3
UK	1	3
Hungary	1	3
Mexico	1	3
Czechia	1	3
Spain	1	3
No answer	5	15
Ever downloaded data?		
Yes	12	36
No	21	64

Category	n	%
Education		
High school	8	24
Some coll.	8	24
Trade/voc.	2	6
Bachelor's	9	27
Master's	4	12
Doctorate	1	3
No answer	1	3
Race		
Asian/Pac. Is.	1	3
Black/Af. Am.	9	27
Multi/Biracial	1	3
White	21	64
Other	1	3
Age		
18-24	16	48
25-34	8	24
35-44	5	15
45-54	3	9
55-64	0	0
65+	1	3

Many participants completed a DSAR for the first time for this study. Of the 33 participants, 12 reported that they had downloaded their data from a platform prior to this study ("ever downloaded data?" in Table 1). In their written responses, nine participants reported doing so out of curiosity, while three did so out of necessity (e.g., because of work). Although Figure 3 shows that most participants did not struggle with the DSAR process, some encountered issues with the delivery timeliness. While we did not formally measure the time to deliver the data export, participants usually had their Google data export by the end of the day. In comparison, Spotify typically took one to two weeks to provide the data.

Google and Facebook, and to a lesser extent Spotify and Amazon, provide controls during the DSAR process, unlike Uber. Facebook users can choose categories of information to include, such as "logged" or "ads" information. Making a DSAR to Facebook or Google requires approximately the same number of steps, but Figure 3 shows participants struggled more with Facebook. This is likely because Facebook's DSAR instructions were briefly outdated since Facebook changed the DSAR process during the study. We had directed participants to Facebook's own instructions, and neither of their two pages[1] were updated to reflect the new procedure.

[1] https://www.facebook.com/help/212802592074644 and https://www.facebook.com/help/contact/180237885820953

Table 2: A description of the structure and characteristics of participants' data exports for the five companies we studied.

Metric	Amazon			Facebook			Google			Spotify			Uber		
# participants	2			9			17			3			7		
# unique keys	749			328			1000			56			99		
# exclusive unique keys	650			149			700			3			0		
Per participant:	min	med	max	min	med	max	min	med	max	min	med	max	min	med	max
# files	27	51.5	76	9	23	56	1	21	53	8	10	10	7	8	10
# unique keys	250	424	598	45	78	188	7	118	545	41	48	55	67	79	99
# directories	5	17	29	4	12	30	2	15	30	1	1	1	4	4	4
Directory depth	-	-	5	-	-	4	-	-	6	-	-	1	-	-	2
Export size (kB)	194	806	1,418	83	976	6,678	5	4,118	38,318	10	260	952	10	1,243	18,966

4.2 Structure of data exports (RQ1)

We received 801 (pseudonymized) files from 33 participants. Table 2 characterizes these files. Spotify and Uber's data exports were the most consistent in the metrics we observed, with two out of the three exports from the former containing files with the same names. The third export was missing the `Identity.json` and `SearchQueries.json` files, although it is unclear why. The identity file in particular contains important data, such as the user's display name and profile photo.

Google, the platform with the most participants, unsurprisingly showed the greatest variation in data export size and complexity, with a median of 15 directories per participant ("# directories" in Table 2). This complexity is understandable since most top-level directories are named after individual Google products (e.g., `Home App/` and `Maps/`). All Google data exports had a `My Activity/` directory, which seemed to be a log of different types of activities for different products, although it is sometimes unclear what these activities are. These activity files, along with history files, were among the largest. To our surprise, these history files were not consistently named across Google participants. For instance, the browsing history was named both `Chrome/History.json` and `Chrome/Browser History.json`.

Participants completed our study from 11 different countries (see Table 1) and had different language preferences (see Appendix C for export complexity across regions). Google was the only platform for which we received filenames in different languages; however, we speculate this is because participants for the other platforms all had set English as their preferred language. Even then, keys and column headers in files with non-English filenames were still in English.

To further understand the complexity of these exports, we counted the number of unique keys within every file in a data export (# unique keys in Table 2). The sample file in Figure 2, for example, would have four unique keys: `my_best_friends`, `first`, `last`, and `handle`. Amazon and Google's largest exports both had a similar number of unique keys (598 and 545, respectively), but the median number of keys per Amazon participant (424) was more than double that of Google (118). This means Amazon has a higher density of unique keys given that the median export size (806 kB) was about one-fifth that of Google's (4,118 kB). Like other metrics, the number of unique keys varied across platforms.

4.3 Participants' goals for their data (RQ2)

Before interacting with their data, we asked participants to think of three to five questions (depending on the size of their exports) about what they wanted to learn from their data. We refer to these as **pre-annotation questions**, After the data annotation exercise (during which participants explored their data, described in detail in Section 4.4), we showed participants their pre-annotation questions and asked them to write a few sentences on whether or not they were able to answer each question; we refer to these as **post-annotation reflections**. While some participants answered with certainty (e.g., "*I did not find the answer in my data,*" P17-Google), others were more vague (e.g., "*there is too much to look into,*" P4-Amazon). We considered pre-annotation questions to be fully answered if the post-annotation reflection contained the answer itself or an affirmative indication that the user found an answer (e.g., "*yes, I got this information,*" P11-Uber). Finally, we asked participants to brainstorm three more questions (**post-annotation questions**) that they still had about their data or the platform after exploring their data export.

In this section, we describe the types of questions participants asked, which of them were answered, and how they evolved with the added context of the data export.

Questions answered by the data export. We first categorize the pre-annotation questions for which participants successfully found answers during the data annotation exercise. Table 3 lists the different types of questions along with the number of participants who asked them, the number of questions of that type, and how many of those were answered.

Closed-ended questions about the existence of records in the export and generic questions about the contents of the files were most commonly answered. In total, 24 participants asked 59 questions about whether a specific record type existed in the data ("specific participant records" in Table 3). These questions were typically short and simple, often in the form of "*Is my location stored in the data?*" (P28-Amazon), al-

Table 3: Key questions participants initially hoped their data download would answer, the number of unique participants (#) who listed that type of question, total questions in that category (?), and the number of questions fully answered (✓) after exploring.

Question Type	#	?	✓	Example Participant Quote
Platform information	**29**	**93**	**59**	
Selling & sharing data	10	17	10	"*Does the company share [its] users' data with other companies?*" (P14-Google)
Storage	10	12	10	"*How long is my data kept for?*" (P33-Google)
Usage (general)	8	10	5	"*Why do you gather my data?*" (P35-Google)
Business practices	7	15	5	"*Does the company use AI to make specialised playlists*" (P4-Spotify)
General data questions	7	10	10	"*What data about me is being collected*" (P6-Google)
Access	7	9	5	"*Who has access to my data?*" (P33-Google).
Privacy & security (general)	6	9	5	"*Is my data safe?*" (P16-Google)
User actions available	4	7	6	"*Can I ask you to delete my data?*" (P15-Google)
Personalization	3	4	3	"*Do they use my position data to show me personalized ads?*" (P23-Google)
Specific participant records	**24**	**59**	**15**	
Location	11	13	11	"*Does Uber track my travel locations?*" (P18-Uber)
Interactions with other users	4	5	3	"*How many people did I engage with?*" (P27-FB)
Nonconsensual data	3	4	3	"*Is the company collecting data it said it wouldn't?*" (P22-Spotify)
Devices	3	3	3	"*Are the devices where I've logged on stored and detailed?*" (P3-Uber)
Browsing/search history	3	3	3	"*Does Amazon know my browser history when suggesting me stuff?*" (P15-Amazon)
Time spent on platform/activities	3	3	2	"*Do they keep track of how much time I spend in the app?*" (P23-FB)
Payment information	3	3	2	"*How are my payment details stored?*" (P28-Amazon)
Other records	15	24	16	"*How many reports or clarifications have I generated in the app?*" (P11-Uber)
Participant information	**13**	**30**	**7**	
Patterns, trends, & hypotheticals	7	9	3	"*What locations have I been to based on my location history, and can I identify any significant travel patterns or places of interest?*" (P13-Google)
First/most of record	6	9	3	"*What was my first friend request I got?*" (P12-FB)
Total of record	5	8	0	"*How many hours have I spent on Facebook in a certain period of time?*" (P27-FB)
Platform's perception of user	3	3	1	"*Compared to other users, am I a good user to the platform?*" (P27-FB)

though some used the term "*tracked*" instead. By far the most requested data type was location, with 11 participants ("location" in Table 3). Several other participants listed concrete data types like age, browsing history, or payment methods, but other participants were vague when referring to data types.

Data exports were effective at answering these questions about specific records. In fact, most of these questions were answered—only seven were not (two each in "location" and "interactions with other users" and one each in "nonconsensual data," "time spent on platform/activities," and "payment information" in Table 3). Most participants noticed that the data type in question was being tracked, but their reactions varied. P1, who completed the study for Google and Facebook, found that both companies had "*a bit more*" location data "*than expected.*" P15-Amazon, on the other hand, felt Amazon "*barely*" knew their "*browser history*."

Participants also asked more general (still often closed-ended) questions about the data. Six participants asked about the security or privacy of their data ("privacy & security (general)" in Table 3), such as P25-Google: "*Is my data handled with caution?*" Ten participants asked how data was stored or for how long it was stored ("storage" in Table 3), while others asked generic questions ("general data questions" in Table 3), like P6-Google: "*What data about me is being collected?*"

Participants who asked generic questions about how much data the platform was storing were able to find answers 10 out of 12 times ("storage" in Table 3). Participants found that they were storing "*a lot*" (P6-Google) of data, "*much more than [they] expected*" (P20-FB). Participants using Uber, Amazon, or Spotify were not quite so surprised, such as P10-Uber, who answered their own question about their Uber trips with "*whole data is here so it's answered.*" Some answers were quite precise ("*signed in data expires after 26 months,*" P34-Google), others less so (data is kept "*forever,*" P33-Google).

Questions unanswered by the data export. We found that questions that required analysis or computation ("patterns, trends, & hypotheticals," "first/most of record," and "total of record" in Table 3), questions about a platform's business practices or technology ("business practices" in Table 3), and questions about specific security practices ("selling & sharing data" in Table 3) were not answered.

Six participants exploring Uber, Facebook, and Amazon asked about the first or earliest instance of a record (e.g., "*first friend,*" P12-FB, or "*oldest purchase,*" P28-Amazon, "first/most of record" in Table 3). These questions were similar to the "specific participant records" in that they were usually short and simple questions with a concrete answer (in this case, a number), but they did require more analysis or synthesis than simply whether a record was present.

Seven participants asked longer and more complex questions ("patterns, trends, & hypotheticals" in Table 3) about their "*communication patterns*" or "*online behavior*" (P13-Google). P24-Uber went a step further: "*Can I use my data to increase the quantity and percentage of discounts I get?*"

Participants received fewer answers to these types of questions. No participant who asked about the total amount of money, time spent, or distance traveled over time got an answer ("total of record" in Table 3). Additionally, only one of P13-Google's questions (their "*most frequently visited website*") was answered. P28-Uber bemoaned the "*lack of summary or any form of analysis,*" which also accounted for the "*disappointing*" lack of answers to P13-Google's questions about their communication patterns and online behaviors. The content of the data exports tended to be mostly raw data—any summaries or syntheses of information were absent. This is likely why most participants were unable to find answers to these questions.

In addition to wondering *what* data platforms stored, participants wanted to know *how* it was used. For example, ten participants wondered if their data was sold to third parties (e.g., "*Does Uber sell client data to telemarketers?*" P24-Uber, "selling & sharing data" in Table 3). Further, 15 questions regarding the platform were about its business dealings ("business practices" in Table 3). P4-Spotify, for example, asked, "*Does the company use AI to make specialised playlists?*"

A little more than half the questions about platforms' usage of data were answered. There were no patterns where some questions were better answered than others ("usage (general)" and "selling & sharing data " in Table 3). In fact, both P25-FB and P9-FB asked if their Facebook data was being sold for the platform's benefit. P25-FB concluded "*Yes, by ad partners*," while P9-FB wrote, "*I don't know more about this point.*" Out of 15 questions about the business side of the platform (excluding those related to selling or sharing data), 10 remained unanswered, with no indication as to why there was no answer ("business practices" in Table 3). GDPR requires platforms to disclose how they use consumer data, but not necessarily within a data export. These unanswered questions could indicate incorrect expectations of what a data export would contain, or a desire to have these exports cover a wider range of topics (e.g., a platform's privacy policy).

New questions. Compared to pre-annotation questions, post-annotation questions (Table 3) focused less on the participant and more on the platform. Only four participants had lingering questions about themselves ("participant-related" in Table 4), which were most similar to the "patterns, trends, & hypotheticals" in Table 3. In fact, some participants carried questions over, such as P26-Uber, who asked about when they used Uber the most both before and after the annotation task. Only seven participants asked about the existence of specific records post-annotation, compared to 24 pre-annotation ("specific participant records" in both Table 3 and Table 4), again indicating that data exports in their current form are well-equipped to address these types of questions.

Questions about "usage," "privacy & security," "storage," and "business practices" were consistent between pre-annotation and post-annotation questions. All participants

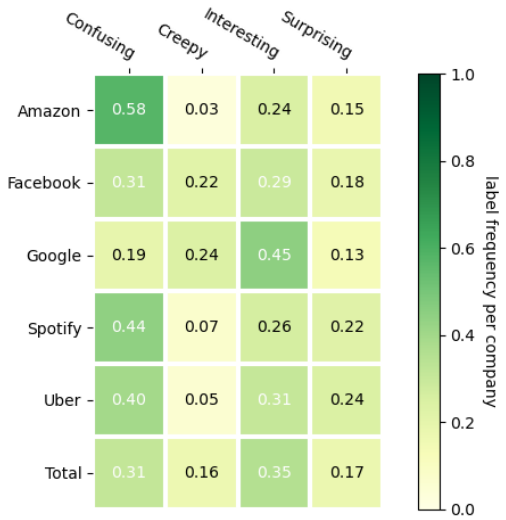


Figure 4: The predominance of the four tags by platform. Percentages are with respect to total annotations per platform.

who had these questions before annotation still asked questions in at least one of these categories post-annotation, although the details sometimes changed. Three participants who asked questions about the business operations of the platform became curious about comparing data across companies, such as P3-FB, who asked, "*How does the data in the EU differ from data in the US?*" This pre- and post-annotation consistency indicates that although current data exports may not be suited to answer questions about how platforms use and manage user data, these questions are important to users.

4.4 Reactions in data annotations (RQ3)

In this section we present findings from the data annotation exercise, the main part of our study. In the exercise, we asked participants to highlight segments of their data that stood out to them, categorize that segment (*confusing*, *creepy*, *interesting*, or *surprising*), and explain why it stood out to them.

Figure 4 shows the frequency with which each of the four labels was used. The most common label was *interesting*, making up almost 40% of the labels applied. However, this finding is influenced by the large number of Google exports we analyzed, of which *interesting* was the most common label. For every other company, the most used label was *confusing*, hovering from around 30% (Facebook) to nearly 60% (Amazon) of labels. Participants did not use the *creepy* label often. It only exceeded 20% in the case of Google and Facebook. For all other companies, its use was below 10% of annotations.

Overall, though, there was no drastic difference in the number of times each label was used. Table 5 suggests the reason—participants often had different reactions to discovering similar or identical pieces of information. Consider two participants for the same platform: P1-Google commented that Google collects "*too much info and they store it for too*

Table 4: After participants explored their data, the questions they still wanted to have answered, including the number of unique participants (**#**) who listed that type of question and total number of questions in that category (**?**).

Question Type	#	?	Example Participant Quote
Platform	**32**	**103**	
Usage	11	18	"*What would they do with information like this?*" (P3-Uber)
Privacy & security (general)	9	11	"*Is it really safe that they have so much information about the user?*" (P11-Uber)
Deleting data & other actions	9	18	"*How can I delete all this info?*" (P9-FB)
Business practices	9	13	"*How is the data used to recommend new music?*" (P4-Spotify)
Storage time period	7	8	"*How long is all this data stored?*" (P7-Google)
Specific participant records	7	9	"*Does Spotify collect banking details?*" (P4-Spotify)
Confusion	6	9	"*What are the strange symbols on some lines?*" (P20-FB)
Missing data	5	8	"*Why isn't there a device login history category?*" (P3-Uber)
Access	5	5	"*Who can access this information besides the user?*" (P11-Google)
How do they know?	3	4	"*How do they have this data?*" (P1-FB)
Others	**6**	**12**	
Participant-related	4	6	"*When do I use Uber [most] often?*" (P26-Uber)
Data annotation platform	2	6	"*What will you do with my data?*" (P33-Google)

long" and P16-Google similarly found the "*very extensive information*" in the Google export to be "*a bit disturbing*." Despite these similar annotations, P1-Google labeled theirs as *creepy* while P16-Google labeled theirs as *interesting*. Thus, we describe high-level themes that crossed label categories.

Surprise the company knows this information. While we never explicitly asked participants about privacy or security concerns before the data annotation task, Table 5 shows themes related to these concerns appeared in all four categories. Participants often questioned why platforms had specific records, how they had them, and how long they planned on storing them. Participants sometimes pointed to specific types of records that should not be collected or stored at all.

Most participants (20) discovered types of records that they had not previously known the platform collected ("tracking, collection, & use" across the four labels in Table 5). These included spending habits, payment methods, emoji use frequency, search history, browsing history, login location, times when activities started, times when activities ended, IP addresses, phone carriers, devices used, car ownership, and locations where they downloaded apps. Participants responded with surprise and confusion. P6-Uber, for example, was not "*sure why [it] is necessary*" to keep their "*bank details in the payment methods file*." The level of detail and specificity of the data often alarmed participants, such as P2-FB, who noted "*Emoji count? Why? I just... Why?*" after discovering a file with counts for every emoji they had ever used on Facebook.

Some participants felt like their preferences had been violated, noting they either opted out of or never knowingly opted into data collection. P20-FB, for example, wondered, "*So does the location pick me up even though my location is not activated???*" (`where_you're_logged_in.json`). P21-Google, plus three other Google participants, realized "*deleting browser history does not mean that [their] data is removed forever.*" P2-FB also felt their preferences were violated when they encountered their off-Facebook activity file:

> "*Words can not really express. Why is Facebook tracking my off Facebook activity? How? Why? Where did I agree to this? How do I opt out? This is by far the most disturbing thing.*"

Sources of confusion and barriers to comprehension. Participants were sometimes unable to interpret the meaning of components within a data export, including file names, keys, and values. Nine participants were confused by the meaning of the values in their data exports ("value meaning" in Table 5). Some of these values were identifiers (e.g., `orderId`, `advertiserName`) and dates (e.g., P13-Google said their `birthday` key was set to 0000), but in most cases our pseudonymization prevented us from investigating further (e.g., `description: *** *** **** ******* **** ..... @`).

Eight participants could not decipher the meaning of some keys within their data exports ("key/column naming" in Table 5). P23-FB, for example, did not "*understand what they mean by 'off Facebook activity'*" in `your_off-facebook_activity.json`. Other keys that caused confusion were `Used For Disbursements` (in `payment_methods-0.csv`, Uber), `PrimeStudentMarkedForGraduation` (in `Subscriptions.PrimeTransition.csv`, Amazon), and `voting_location` (in `voting_location.json`, Facebook). No participant mentioned finding the meaning of these terms in their data export.

Some participants (14) did not find data they were expecting ("missing & incomplete data" in in Table 5). Many keys had blank or null values for no clear reason (e.g., P8-Spotify said `episode` was set to `null` in `Playlist1.json`). Participants could not understand why some records had "*such careful tracking of events*" (P18-Uber) and nothing for data participants expected the platforms to have (such as P18-Uber's order history for Uber Eats). Two participants mentioned seeing incomplete Spotify listening history. Here, the missing data was expected as we asked participants to exclude streaming history from Spotify DSARs (see Section 3.1).

Table 5: Clustering of participants' annotations, with the number (#) of unique participants creating each, whether the subject of the annotation is the platform (🏭) or participant (👤), and an example.

	Annotation Theme	#	Example Annotation
	Confusing	**30**	
🏭	Missing & incomplete data	14	P16-Google opened their YouTube search history file and noted the history "*is incomplete, I don't know what is the reason, because I watched more materials.*"
🏭	Value meaning	9	P32-Google found a list of Chrome extension ids and noted "*I don't understand what is written there.*"
🏭	Key/column naming	8	P2-FB annotated a line in `your_record_details.json` with "*What the hell is a checkpoint?*"
🏭	Tracking, collection, & use	9	P4-Spotify annotated `Inferences.json` with "*Looked at all files and I have no clue what they are for. This doesn't sound like something connected to music at all.*"
🏭	Deleted records still exist	3	P21-Google annotated the `Search/My Activity.json` file with "*Google keeps my data from a very long time ago even though I sometimes delete browser history of all time on all my devices.*"
🏭	Inferred data	3	P23-FB found a list of topics Facebook thinks are interesting to them in `feed.json` and noted "*I don't know why they think this a interesting topic for me.*"
🏭	Befuddlement	3	P9-FB annotated 4 lines in `advertisers_you've_interacted_with.json` with "*??????*"
	Creepy	**23**	
🏭	Tracking, collection, & use	18	P20-Google has "*no idea why*" the Google Play Store order history would be "*showing locations.*"
🏭	Old/stale data storage	5	P6-Google noted "*Wow that was a long time ago - 6 years of activity tracked?*" when exploring their YouTube activity.
🏭	Preferences violated	6	P25-Amazon noted in `consents.json` that they "*do not remember consenting to ad partners.*"
🏭	Nonsensical/incorrect data	4	P12-FB "*I don't even [know] who those people are*" listed in `people_who_followed_you.json`.
🏭	How to stop collection	3	P9-FB wondered "*how can I delete this?*" in the `recognized_devices.json` file.
	Interesting	**33**	
🏭	Tracking, collection, & use	13	P25-Google noted "*I did not know that my steps are counted*" in `derived_com.google.step_count.delta_com.google(2).json`.
🏭	Platform inner-workings	10	P1-FB noted "*I didn't know this option existed*" in `ad_preferences.json`.
👤	Self-discovery	9	"*It helps me visualize when I started on Uber Eats. It has been longer than I thought*" (P18-Uber in `eats_restaurants-0.csv`).
👤	Memory & recognition	7	"*It reminds of the beautiful pictures I captured back then*" (P27-FB, `eats_restaurants-0.csv`).
🏭	Against data collection	8	P3-FB noted a file with administrative records "*seems excessive to store.*"
🏭	Missing & incomplete data	8	"*Entries are less detailed than anticipated*" (P18-Uber, `eats_restaurants-0.csv`).
🏭	Confusion	6	P13-Google noted in the video searches activity file that "*This could have the real title to be more easy to identify.*"
👤	Useful data	5	"*Detailed information about used phones in your account. Very useful information*" (P16-Google, `Devices.json`).
	Surprising	**29**	
🏭	Tracking, collection, & use	14	"*They have my address and I don't like it*" (P1-Google, `Addresses and more.json`).
🏭	Missing & incomplete data	7	P11-Uber noted that "*In such a complete report, they should have*" IP addresses in `rider_app_analytics-0.csv`.
🏭	Nonsensical/incorrect data	6	"*I'm not sure I downloaded these apps*" (P1-Google, `Google Play Store/Library.json`).
👤	Self-discovery	5	"*I listen to quite a lot of rap*" (P19-Google, playlist video file).
👤	Memory & recognition	3	"*This is one of the games I played with some friends about 6 years ago*" (P12-FB, `instant_games.json`).

Sometimes when data was present, participants noted it was wrong or nonsensical ("nonsensical/incorrect data" in Table 5). Participants noticed some locations listed in their data had no relation to them. P13-Google, for example, noted they "*don't even know*" the address listed under their saved locations, and asked, "*How can it be my home address?*" In other cases, participants could not remember taking actions recorded in the exports. P23-FB, for example, did "*not remember these people nor saving any products on Facebook*" in the `your_saved_items.json` file. Many of these notes were framed as questions, indicating participants are lacking key information to fully understand the data in their exports.

Self-discovery, memories, and nostalgia. Some data exports helped participants learn about themselves or reflect on important memories. For example, 10 participants enjoyed making discoveries about the advertisers and interest categories associated with them ("platform inner-workings" in Table 5). P28-Amazon "*had no idea these many brands bought audiences with [them]. Some of these [they]'ve never heard of*" while exploring `Advertising.AdvertiserAudiences.csv`. Sometimes these findings were amusing, like for P22-Spotify, who noted "*lol, I guess I'm not very profitable. I like that*" on the inferences (here, likely age group) the platform had made about them.

In other cases, participants learned about their own behaviors and attitudes ("self-discovery" in the "interesting" and "surprising" categories in Table 5). For example, P19-Google noticed they "*watch quite a lot of educational content*" and "*listen to a lot of rap*" on YouTube. Analyzing data longitudinally, P11-Uber was "*surprised to know the waiting times and [idle times] have increased over the years*" when exploring

their Uber trips data. These examples show participants found useful—or at least interesting—information.

In addition to learning new information, eight participants recalled specific events or people after seeing a related record in their data ("memory & recognition" in the "interesting" and "surprising" categories in Table 5). While some of these annotations were neutral ("*I remember visiting [this page] on my phone,*" P33-Google), others had emotion. P27-FB, for example, "*loved*" the Facebook story they found in `stories/story_reactions.json`. P12-FB found a "*childhood friend of [theirs] which [they] put as brother in FB 'cause [they] were very close*" while looking through `profile_information.json`. Not all of P12-FB's memories were positive ones: they later found people "*deleted from their friend list*" whose "*name reminds [them]*" why they were removed.

4.5 Deleting data and other actions (RQ4)

In their pre-annotation questions, annotations, post-annotation reflections, and post-annotation questions, participants mentioned wanting to take action. The most commonly requested action was deleting some, not all, of their data. While only two pre-annotation questions about data deletion were left unanswered, sometimes answers from separate participants contradicted each other. Two Google participants reached separate conclusions when wondering if consumers could delete their data—one noted it "*doesn't seem*" like a possibility (P13-Google), while the other answered "*absolutely*" (P34-Google). If there were instructions for making data deletion requests in P13-Google's data export, they could not find them. Some participants, such as P2-FB, expressed skepticism about whether platforms would follow through:

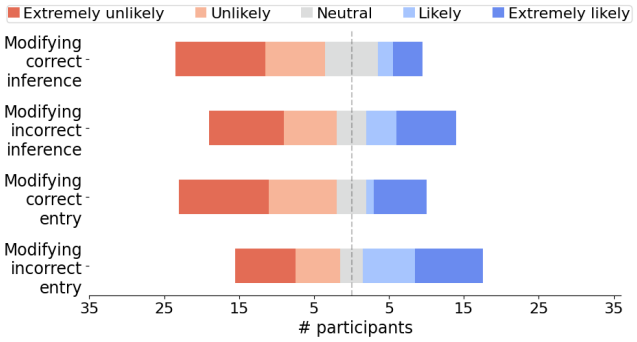
> "*If they do delete, how much of this will be actually deleted and not just kept anonymously?*"

Participants sometimes thought data deletion was something platforms should implement automatically. P3-FB thought their contacts "*should be wiped after some time maybe*" and P1-Google thought Chrome's browser history was stored "*for too long.*" P21-Google suggested a solution:

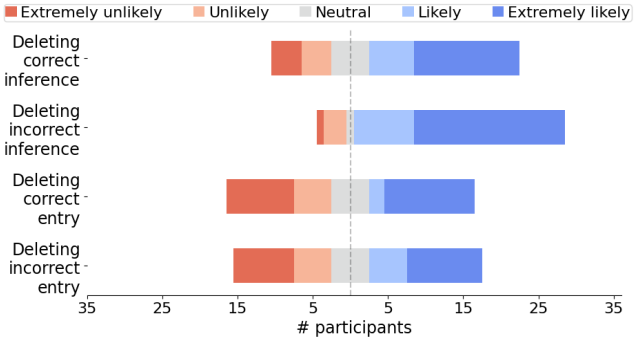
> "*I think there should be a way to remove my history across all Google platforms by just deleting on the Chrome browser. Also, there should be a choice by myself as a user as to which data about me I want kept by Google, thus being in control of my data.*"

Participants were also concerned with future data collection and usage. For example P9-FB asked, "*How can I withdraw my consent?*" from the apps connected to Facebook listed in `connected_apps_and_websites.json`.

Our findings from the post-annotation survey confirm that participants were interested in data deletion. In the survey, we asked participants how likely they would be to modify or delete the following: an outdated home address they entered



(a) Likelihood to **modify** data.



(b) Likelihood to **delete** data.

Figure 5: Participants' reported likelihood to modify or delete data based on correctness and whether it was either inferred by the platform or entered by the user.

into the platform, an accurate home address they entered, an incorrect political party inference by the platform, and a correct political party inference. Figure 5 shows that participants were more interested in data deletion than modification. Figure 5b shows that participants were less likely to delete data they entered themselves compared to data inferred by the platform. Some said their home address was "*shared public information*" (P18-Uber) and therefore not a privacy violation. Others believed the information was relevant to the platform (e.g., "*this data seems to be important for Uber,*" P10-Uber), but this attitude was often company specific—meaning they may be content with Uber having their information, but "*Spotify shouldn't have [their] address*" (P38-Spotify).

5 Discussion

In this study, we asked participants to annotate and answer questions about data exports resulting from DSARs to five different companies (Amazon, Facebook, Google, Spotify, or Uber). We then analyzed their survey responses, annotations, and the exports themselves in order to characterize the content and variability of these exports, as well as how participants wanted to use them and which content they found galvanizing.

Overall, participants were surprised to find specific data (e.g., search history) being collected or retained by platforms.

They mentioned different facets of tracking found in several files within the exports. Participants were able to use their data exports to answer simple questions, but also had many more complex questions that data exports are not currently equipped to answer. Participants also expressed interest in integrating actions (e.g., deleting or correcting data) directly into the workflow of viewing their data exports.

Next, we discuss the implications of our findings as they relate to the GDPR [21] and other laws with similar provisions. We then provide suggestions for developing new tools for consumers to explore data exports.

5.1 Policy shortcomings and recommendations

Reduce time to deliver. Article 12 of the GDPR requires that platforms respond to DSARs "without undue delay" (30 days) [21]. While all five of the platforms we investigated responded to participants within 30 days of the DSAR, the delay in these responses varied between hours (Google), hours to days (Facebook, Uber), and days to weeks (Spotify, Amazon). After delays of days or weeks, some participants missed the email notification that the data export was available. When they found it, the link had expired, requiring these participants to send another DSAR. **Requiring companies to reduce the time consumers have to wait for DSAR responses to at most 15 days would likely improve accessibility**.

Provide data summaries and definitions. In Section 4.2, we showed that many data exports are not concise, an explicit requirement of Article 12 of the GDPR (data must be "concise, transparent, intelligible and easily accessible"). Some exports are large in size (in Google's case, sometimes over 30 MB worth of text-based files) and have several labels for different records, demonstrated by the dozens to hundreds of unique keys in JSON files. The data is not always transparent or intelligible; participants were often confused by JSON keys (e.g., `Horizontal Accuracy`) and their associated values and did not understand the meaning of specific records, and sometimes even entire files. **Companies should be required to provide definitions for every file, term, record, and key in the consumer's preferred language, as a significant step in transparency and intelligibility of data exports**. We recommend policymakers incorporate the "data dictionary"—a system for providing information (e.g., a variable's possible values, human-readable names, variable definitions) about variables in a dataset [9, 12]—into legislation to standardize how this information is communicated. Striking a balance between concision and clarity is more complex since platforms are obliged to send consumers all the requested data, even if it spans many gigabytes and hundreds of files. Therefore, in addition to the entire data export, **platforms should be required to create record summaries**, which could include the total number of records of each type or the earliest date a specific record can be found in the export.

Justify data retention. The GDPR, under Article 15, also guarantees that data subjects receive "confirmation as to whether or not personal data concerning him or her are being processed" and if so, the "purposes of the processing" [21]. Table 4, however, shows that participants' primary remaining questions after exploring their data exports were precisely about how their data was being used. Our results suggest these questions are both important to consumers and difficult to answer with current data exports. Along with providing definitions for records, **platforms should be required to provide a clear description of how the records are used alongside the records themselves**.

5.2 Designing a data-exploration tool

We expect platforms and tools built for personal data exploration to continue growing in popularity [46, 53, 55]. Based on participants' reactions, we suggest that future designers and developers focus on the following three areas. While our data tool (Section 3.3) was a means to an end by helping participants explore and annotate their data (and therefore not the focus of our study), we also reflect on how our tool could be adapted to embody these three principles.

Data visualization and interaction without truncation. Participants' questions often involved discovering patterns in their usage of the platforms. P28-Uber, for example, asked: "*Can you show a summary of my data or perhaps graphs and analysis to understand how I use Uber?*" We suggest future tools do just that—provide ways of interacting and visualizing personal data, which many existing tools already do. However, many participants were confused and disappointed by missing data in their exports. While this is likely due to the DSARs they submitted excluding specific categories of data, it shows that participants want every piece of data to be accessible to them. Developers of these tools will need to come up with creative ways of showing aggregated data while also letting participants delve into specific records in the export.

While our tool did not truncate any records within the files it parsed, we did not provide any visualizations or data syntheses. As a result, the experience was overwhelming for participants. In an early prototype of the tool, we abbreviated some of the data by limiting lists within JSON objects to three items and CSV tables to ten items. While this was effective at reducing files that sometimes were split into thousands of pages to fewer than ten, the pilot participants often mentioned wanting access to all of their data. We recommend developers explore the trade-offs between abbreviation and comprehensiveness, but ultimately not sacrifice the latter.

Designing for meaningful interaction. Synthesizing and summarizing data to show users the broader picture is important, but it is not the only way our participants wanted

to interact with their data. Our participants had strong emotional reactions to individual records that triggered memories and recognition. P19-Google, for example, was able to find a record for an important milestone: the "*moment when [they] started listening to audiobooks not in [their] native language*." Other individual data points triggered surprise or alarm by showing participants that their data was collected or used in unexpected, sometimes upsetting ways. Advanced exploration tools could add features designed to surface these highly salient individual events or records (positive or negative), perhaps by taking advantage of advances in machine learning to answer complex questions or identify likely candidate events. Doing so could enable data exports to provide not just simple transparency, but also understanding [8].

Our tool did not save participants' annotations on the client side once the study was finalized, but future tools could implement bookmarking for participants to retain records of interest. While we did not ask if participants would be interested in such a feature, allowing them to save records might be a straightforward, but meaningful, interaction.

Enabling action in-band. Participants expressed strong interest in taking action in response to things they found in their data exports, including deleting existing data and withdrawing their consent for a platform to collect similar data in the future. Such a feature was entirely absent in our tool as its purpose was exploratory. The right to delete personal data ("erasure," Article 17) and to withdraw consent (Article 7) are both granted by the GDPR [21]. Currently, consumers who want to take these actions after reviewing their data exports must visit the platform and find potentially difficult-to-access options and settings. A better interaction mode would enable consumers to take actions directly, in-band, as soon as they decide the actions are needed. Tools should build in mechanisms for taking these actions. Of course, this will depend on the ability for independent data-exploration tools to communicate these desired actions on behalf of users via an API or other automated submission, as has been done for the right to access [34]. It is unclear whether platforms have incentives to support this feature absent further regulation.

6 Conclusion

Recent legislation in several countries has made it easier for consumers to access their personal data collected by companies. However, prior work has shown that consumers are often unsure of how to make sense of the data they receive, are overwhelmed by the amount, and sometimes cannot even open the data files [62]. In order to develop better ways for consumers to interface with their personal data, either directly through the files they receive from a company or through a tool built for users to interact with their data with more ease, we set out to understand what participants want to learn from their data and what their primary reactions are when exploring it. To answer these questions we designed a study where consumers of five different platforms (Amazon, Facebook, Google, Spotify, and Uber) interacted with their data exports firsthand to share their questions and reactions to seeing their data. Part of this study involved the development of a data annotation tool where participants were able to apply labels (confusing, creepy, interesting, or surprising) and attach notes to segments of their data exports that stood out to them.

Participants were primarily interested in learning about their usage of the platform, what data the platform collects, and how the platform uses that data. While most participants were able to find the types of data the platform collects in their data exports, questions about the platforms' use of the data and participants' usage of the platform were often left unanswered after exploring the data exports. Participants had a wide range of reactions to exploring their data. For some, fond memories were triggered when they encountered a record. Other times, participants were surprised to learn about the scale and detail of the personal data held by the platform. As a result, participants were interested in learning how to make requests to have parts of their data deleted.

Based on our findings, we suggest that companies improve the clarity of their data exports by providing easily accessible definitions for files and terms. For designers of tools for personal data exploration, we provide a characterization of different types of questions participants wanted answered as a source for building relevant data experiences for consumers.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. CNS-2047827, CNS-2149680, and CNS-2151290. We thank Hayley Szymanek and Sophie Veys for additional assistance with our web app's implementation. We also thank Grant Nakanishi and Maya Thumpasery for assisting with qualitative coding.

References

[1] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location Has Been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *Proc. CHI*, 2015.

[2] David Alpert. Beyond Request-and-Response: Why Data Access Will Be Insufficient to Tame Big Tech. *Columbia Law Review*, 120:1215–1254, 2020.

[3] Julio Angulo, Simone Fischer-Hübner, Tobias Pulls, and Erik Wästlund. Usable Transparency with the Data Track: A Tool for Visualizing Data Disclosures. In *Proc.. CHI EA*, 2015.

[4] Patricia Arias-Cabarcos, Saina Khalili, and Thorsten Strufe. 'Surprised, Shocked, Worried': User Reactions to Facebook Data Collection from Third Parties. *PoPETs*, 2023(1):384–399, 2023.

[5] Theo Bertram et al. Five Years of the Right to Be Forgotten. In *Proc. CCS*, 2019.

[6] Christoph Bier, Kay Kühne, and Jürgen Beyerer. PrivacyInsight: The Next Generation Privacy Dashboard. In *Proc. APF*, 2016.

[7] Coline Boniface, Imane Fouad, Nataliia Bielova, Cédric Lauradoux, and Cristiana Santos. Security Analysis of Subject Access Request Procedures. In *Proc. APF*, 2019.

[8] Will Brackenbury, Rui Liu, Mainack Mondal, Aaron J. Elmore, Blase Ur, Kyle Chard, and Michael J. Franklin. Draining the Data Swamp: A Similarity-Based Approach. In *Pro. HILDA*, 2018.

[9] Erin M. Buchanan, Sarah E. Crain, Ari L. Cunningham, Hannah R. Johnson, Hannah Stash, Marietta Papadatou-Pastou, Peder M. Isager, Rickard Carlsson, and Balazs Aczel. Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set. *AMPPS*, 4(1), 2021.

[10] Luca Bufalieri, Massimo La Morgia, Alessandro Mei, and Julinda Stefa. GDPR: When the Right to Access Personal Data Becomes a Threat. In *Proc. ICWS*, 2020.

[11] California State Legislature. California Consumer Privacy Act, 2018.

[12] Center for Open Science. OSF Support: How to Make a Data Dictionary, 2024. https://help.osf.io/article/217-how-to-make-a-data-dictionary.

[13] Aloni Cohen, Adam Smith, Marika Swanberg, and Prashant Nalini Vasudevan. Control, Confidentiality, and the Right to be Forgotten. In *Proc. CCS*, 2023.

[14] Jessica Colnago, Lorrie Cranor, and Alessandro Acquisti. Is There a Reverse Privacy Paradox? An Exploratory Analysis of Gaps Between Privacy Perspectives and Privacy-Seeking Behaviors. *PoPETs*, 2023(1):455–476, 2023.

[15] Personal Information Protection Commission. Amended Act on the Protection of Personal Information. 2020.

[16] Data Privacy Manager. What is a Data Subject Access Request (DSAR), 2021. https://dataprivacymanager.net/what-is-data-subject-access-request-dsar/.

[17] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *PoPETs*, 2015(1):92–112, 2015.

[18] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services. *CLSR*, 34(2):193–203, 2018.

[19] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proc. NDSS*, 2019.

[20] Serge Egelman, Adrienne Porter Felt, and David Wagner. Choice Architecture and Smartphone Privacy: There's a Price for That. In *Proc. WEIS*. 2012.

[21] European Parliament and the Council of the European Union. General Data Protection Regulation, 2016.

[22] Florian M. Farke, David G. Balash, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. Are Privacy Dashboards Good for End Users? Evaluating User Perceptions and Reactions to Google's My Activity. In *Proc. USENIX Security*, 2021.

[23] Federal Privacy Council. Fair Information Practice Principles (FIPPs). https://www.fpc.gov/resources/fipps/.

[24] Mafalda Ferreira, Tiago Brito, José Fragoso Santos, and Nuno Santos. RuleKeeper: GDPR-Aware Personal Data Compliance for Web Frameworks. In *Proc. IEEE S&P*, 2023.

[25] Simone Fischer-Hübner, Julio Angulo, Farzaneh Karegar, and Tobias Pulls. Transparency, Privacy and Trust - Technology for Tracking and Controlling My Data Disclosures: Does This Work? In *Proc. IFIPTM*, 2016.

[26] General Assembly of the State of Colorado. Colorado Privacy Act. 2021.

[27] Ben Gerber and Organisation for Economic Co-operation and Development. OECD Privacy Principles, 2010. http://oecdprivacy.org/.

[28] Thomas Groß. Validity and Reliability of the Scale Internet Users' Information Privacy Concerns (IUIPC). *PoPETs*, 2021(2):235–258, 2021.

[29] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An Empirical

Analysis of Data Deletion and Opt-Out Choices on 150 Websites. In *Proc. SOUPS*, 2019.

[30] Adamu Adamu Habu and Tristan Henderson. Data Subject Rights as a Research Methodology: A Systematic Literature Review. *JRT*, 16, 2023.

[31] Miró Khalifa. Data Access Automation at Scale. In *Proc. PEPR*, 2023.

[32] Jacob Leon Kröger, Jens Lindemann, and Dominik Herrmann. How Do App Vendors Respond to Subject Access Requests? A Longitudinal Privacy Study on iOS and Android Apps. In *Proc. ARES*, 2020.

[33] Lin Kyi, Sushil Ammanaghatta Shivakumar, Cristiana Teixeira Santos, Franziska Roesner, Frederike Zufall, and Asia J. Biega. Investigating Deceptive Design in GDPR's Legitimate Interest. In *Proc. CHI*, 2023.

[34] Nicola Leschke, Florian Kirsten, Frank Pallas, and Elias Grünewald. Streamlining Personal Data Access Requests: From Obstructive Procedures to Automated Web Workflows. In *Proc. ICWE*, 2023.

[35] Tina Marjanov, Maria Konstantinou, Magdalena Jóźwiak, and Dayana Spagnuelo. Data Security on the Ground: Investigating Technical and Legal Requirements Under the GDPR. *PoPETs*, 2023(3):405–417, 2023.

[36] Mariano Di Martino, Pieter Robyns, Winnie Weyts, Peter Quax, Wim Lamotte, and Ken Andries. Personal Information Leakage by Abusing the GDPR 'Right of Access'. In *Proc. SOUPS*, 2019.

[37] Aleecia M. McDonald and Lorrie Faith Cranor. Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising. In *Proc. TRPC*, 2010.

[38] Pankaj Mohapatra. Building a Complete Export Ecosystem-From DSAR Automation to Privacy Center. In *Proc. PEPR*, 2023.

[39] Trung Tin Nguyen, Michael Backes, Ninja Marnau, and Ben Stock. Share First, Ask Later (or Never?) Studying Violations of GDPR's Explicit Consent in Android Apps. In *Proc. USENIX Security*, 2021.

[40] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. On The Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies. In *Proc. ConPro*, 2019.

[41] Marta Piekarska, Dominik Strohmeier, Yun Zhou, and Alexander Raake. Because We Care: Privacy Dashboard on FirefoxOS. In *Proc. W2SP*, 2015.

[42] Jon Porter. GDPR Makes It Easier to Get Your Data, but That Doesn't Mean You'll Understand It. *The Verge*, January 2019.

[43] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. In *Proc. CHI EA*, 2022.

[44] Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Liu, Aleecia McDonald, Thomas B. Norton, and Rohan Ramanath. Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding. *Berkeley Technology Law Journal*, 30(1):1–88, 2015.

[45] Nathan Reitinger, Bruce Wen, Michelle L. Mazurek, and Blase Ur. Analysis of Google Ads Settings Over Time: Updated, Individualized, Accurate, and Filtered. In *Proc. WPES*, 2023.

[46] Rick Rieta and Tony Bui. Festify. https://salty-beach-42139.herokuapp.com/.

[47] Ryan Rix. Securing and Standardizing Data Rights Requests with a Data Rights Protocol. In *Proc. PEPR*, 2023.

[48] Marlene Saemann, Daniel Theis, Tobias Urban, and Martin Degeling. Investigating GDPR Fines in the Light of Data Flows. *PoPETs*, 2022(4):314–331, 2022.

[49] Marija Schufrin, Steven Lamarr Reynolds, Arjan Kuijper, and Jorn Kohlhammer. A Visualization Interface to Improve the Transparency of Collected Personal Data on the Internet. *IEEE TVCG*, 27(2):1840–1849, 2021.

[50] Supreeth Shastri, Vinay Banakar, Melissa Wasserman, Arun Kumar, and Vijay Chidambaram. Understanding and Benchmarking the Impact of GDPR on Database Systems. *Proc. VLDB*, 13(7):1064–1077, 2020.

[51] Supreeth Shastri, Melissa Wasserman, and Vijay Chidambaram. The Seven Sins of Personal-Data Processing Systems under GDPR. In *Proc. HotCloud*, 2019.

[52] Keith Spiller. Experiences of Accessing CCTV Data: The Urban Topologies of Subject Access Requests. *Urban Studies*, 53(13):2885–2900, 2016.

[53] Spotify. 2023 Wrapped on Spotify, 2023. https://www.spotify.com/us/wrapped/.

[54] State of Connecticut General Assembly. Connecticut Data Privacy Act. 2022.

[55] Dora Szucs and Krisztina Szucs. Tinder Insights. https://tinderinsights.com.

[56] Jan Tolsdorf, Florian Dehling, and Luigi Lo Iacono. Data Cart – Designing a Tool for the GDPR-Compliant Handling of Personal Data by Employees. *Behaviour & Information Technology*, 41(10):2084–2119, 2022.

[57] Jan Tolsdorf, Michael Fischer, and Luigi Lo Iacono. A Case Study on the Implementation of the Right of Access in Privacy Dashboards. In *Proc. APF*, 2021.

[58] United States Department of Justice. Privacy Act of 1974. https://www.justice.gov/archives/opcl/conditions-disclosure-third-parties.

[59] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. A Study on Subject Data Access in Online Advertising After the GDPR. In *Proc. DPM*, 2019.

[60] Utah State Legislature. Utah Consumer Privacy Act. 2022.

[61] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proc. CCS*, 2019.

[62] Sophie Veys, Daniel Serrano, Madison Stamos, Margot Herman, Nathan Reitinger, Michelle L. Mazurek, and Blase Ur. Pursuing Usable and Useful Data Downloads under GDPR/CCPA Access Rights via Co-Design. In *Proc. SOUPS*, 2021.

[63] Lun Wang, Usmann Khan, Joseph Near, Qi Pang, Jithendaraa Subramanian, Neel Somani, Peng Gao, Andrew Low, and Dawn Song. PrivGuard: Privacy Regulation Compliance Made Easier. In *Proc. USENIX Security*, 2022.

[64] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L. Mazurek, and Blase Ur. What Twitter Knows: Characterizing Ad Targeting Practices, User Perceptions, and Ad Explanations Through Users' Own Twitter Data. In *Proc. USENIX Security*, 2020.

[65] Savvas Zannettou, Olivia-Nemes Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P. Gummadi, Elissa M. Redmiles, and Franziska Roesner. Leveraging Rights of Data Subjects for Social Media Analysis: Studying TikTok via Data Donations. arXiv:2301.04945, 2023.

A Data-Annotation Platform Details

Figures 6–7 are the first two of five slides participants see when they open the annotation tool. The remaining three are additional instructions for using the tool, similar to Figure 7.

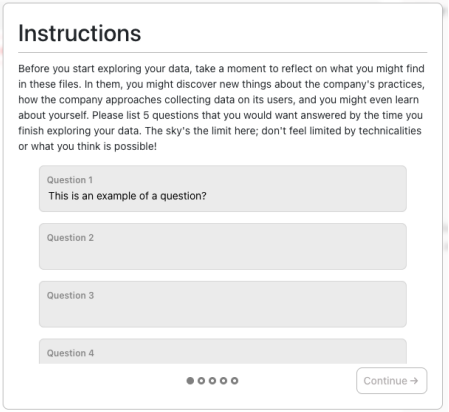


Figure 6: Instructional slide prompting the participant to come up with five questions about their data. The slide post-annotation task that asks for more questions is nearly identical.

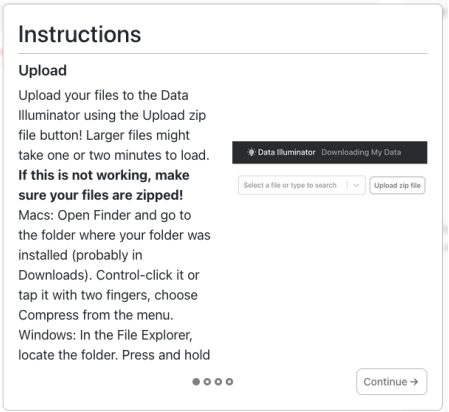


Figure 7: Instructions for uploading the data export.

Participants see three slides before final submission. The first is similar to Figure 6, except it prompts the participant to reflect on the original questions instead of asking new ones:

> Before you started the study, you proposed the following 5 questions you'd want answered by now. For each of them, reflect on the extent to which you have an answer after exploring your data and how you feel about the answer or lack thereof.

The following slide prompts participants to come up with new questions and is even more similar to Figure 6. It reads:

> When you submit your files, we don't receive any of the values of your raw data. For each file, we receive two anonymized versions: one anonymized through hashing and one through direct character replacement.
>
> Hashing replaces every value with a randomized string that allows us to see how often a value appears in your files but not the

value itself. For example, if "USA" were hashed to "89ac13af" we would be able to see how many times "89ac13af" appeared in your files but not know that it corresponds to "USA."

Structural anonymization replaces every alphabetical character (A-Z or a-z) with * and every number with #. The email address abc@xyz.com would become ***@***.*** and the phone number 555-123-4567 would become ###-###-####.

All of your data is stored on your browser until you hit the submit button, and anonymization happens before your files are sent to us. If you have any questions, please reach out to [*contact address*] or message us on Prolific. All set? Submit below!

B Main Study Survey Instrument

[*Consent agreement*]
- Did you receive an email confirmation from [*company*] that your data was ready to be downloaded? • Yes • No, but I've requested my data • No, and I no longer wish to participate in the study
- Have you downloaded a zip file containing your data from [*company*]? Please do so before continuing with the survey. • Yes • No
- What was the subject line of the email that provided access to your data?
[*Participant is then directed to the data annotation tool*]
- Is this the first time you've taken this survey? • Yes • No, I've taken this survey using data from another company before
- Do you think you will download your data from [*company*] in the future? Explain why or why not.
- Before using the [*tool*], did you attempt to view the content of the file(s) downloaded from [*company*] for this study on your own computer? • Yes • No
- Had you downloaded your data from any company or service online prior to participating in this study? • Yes • No
- [*If yes*] From which companies had you downloaded your data prior to this study?
- Why did you download your data from these companies prior to this study?
- Do you think you will download your data from any other company besides [*company*] in the future? • Yes • No
- Why?
- As part of your task while using the [*tool*], you were asked to explore files from your data downloaded from [*company*]. The following questions ask about your process of exploration and discovery when interacting with your data.
- What two strategies did you find most useful to discover new or interesting information from your data? Select at most two. • Randomly picking files from the list of files • Built in search feature of the [*tool*] • Searching for a specific file or folder name • Using your browser's website search feature to search for keywords (e.g., Ctrl-f) • Other
- When exploring your data, which of the following were you interested in learning about or finding? Select all that apply. • Specific records from [*company*], such as your date of birth or your email address, that you know exist • Insights into how you use the service provided by [*company*], such as frequent login times • A record of a specific event or piece of information that might be in your data (e.g., a purchase or post made on a specific date) • Data you may find discomforting for [*company*] to have on you • Inferences about you or advertising targeted to you by [*company*] • Other
- Now we'll ask about your experience exploring data within a file. As you might remember, data within each file was organized using colons and indentation. Suppose the [*tool*] showed the following for a file your friend uploaded.
[*The survey shows a 12-line snippet of a JSON file with a name, an email, a list of cities/ZIP codes where logins occurred, and a list of first and last names of friends.*]
- Describe the content of this file.
- How confident are you in your explanation above? • Not confident at all • Slightly confident • Somewhat confident • Fairly confident • Completely confident
- Suppose that in the future a non-profit focused on internet literacy builds a version of the [*tool*] with several new features focused on visualizing and interacting with your data in more meaningful ways in order to help you understand your data.
- What are some reasons you would use the future [*tool*] when it's released?
- What would make you hesitant to use the future [*tool*] when it's released?
- In recent years Europe and some states in the US have passed legislation to protect and promote data privacy and data access. The ability to download your data from [*company*] or other similar companies is a consequence of such legislation. In addition to giving you access to your data, this legislation also entitles you to request [*company*] or other similar companies to delete any data they've collected from you.
- How these requests are made is a platform dependent decision. For the purposes of this study, assume that you would need to contact [*company*] by email to enact your data rights.
- Suppose while you were browsing your [*company*] data you find that they have a preferred political party on file, but you don't remember ever posting or entering information about this. How likely would you be to request this information be deleted if…
  - the information was correct: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
  - the information was incorrect: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
- Explain your reasoning for your answers above.
- Would your answers change if this were about a platform other than [*company*]? • Yes • No
- Why would your actions change if the company in question wasn't [*company*]?
- Under the same political party scenario (where [*company*] has a preferred political party on file for you, but you don't remember ever posting or entering information about this), how likely would you be to request this information be modified (not deleted) if…
  - the information was correct: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
  - the information was incorrect: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
- Explain your reasoning for your answers above.
- Would your answers change if this were about a platform other than [*company*]? • Yes • No
- Why would your actions change if the company in question wasn't [*company*]?
- Now suppose you come across the home address you entered when you first created your [*company*] account. How likely would you be to request this information be deleted if…
  - the information was correct: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
  - the information was incorrect: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
- Explain your reasoning for your answers above.
- Would your answers change if this were about a platform other than [*company*]? • Yes • No
- Why would your actions change if the company in question wasn't [*company*]?
- Under the same home address scenario, how likely would you be to request this information be modified (not deleted) if…
  - the information was correct: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
  - the information was incorrect: • Extremely unlikely • Unlikely • Neutral • Likely • Extremely likely
- Explain your reasoning for your answers above.
- Would your answers change if this were about a platform other than [*company*]? • Yes • No
- Why would your actions change if the company in question wasn't [*company*]?
[*IUIPC-8 questionnaire* [28]]
- What is your gender? • Female • Male • Non-binary • Prefer not to answer • Prefer to self-describe
- What is your age? • 18-24 • 25-34 • 45-54 • 55-64 • 65 or older • Prefer not to answer
- What is the highest degree or level of school you have completed? • Some high school • High school • Some college • Trade, technical, or vocational training • Associate's degree • Bachelor's degree • Master's degree • Professional degree • Doctorate • Prefer not to answer
- Which of the following best describes your educational background or job field? • I have an education in, or work in, the field of computer science, engineering, or IT • I do not have an education in, or work in, the field of computer science, engineering or IT • Prefer not to answer
- Which of the following best describes your race? Select all that apply. • Asian or Pacific Islander • Black or African American • Native American or Alaskan Native • White or Caucasian • Multiracial or Biracial • A race not listed here • Prefer not to answer
- Which of the following best describes your ethnicity? • Hispanic or Latino • Not Hispanic or Latino • Prefer not to answer
- In what country are you located?

C Export characteristics by participant region

Table 6: The structure and characteristics of participants' data exports by region. Five participants did not disclose their location, while "other countries" are the U.S., U.K., and Mexico. The complexity of data exports varied across regions.

Metric	**Continental Europe**			**South Africa**			**Other countries**		
# participants	14			9			3		
# unique keys	1333			480			858		
# exclusive unique keys	872			304			851		
Per participant:	min	med	max	min	med	max	min	med	max
# files	7	26	84	4	11	43	9	31	53
# unique keys	45	78	188	41	48	55	67	79	99
Export size (kB)	10	987	38,319	5	6,724	18,966	858	5,386	7,625