Italian Sociological Review
ISSN 2239-8589
DOI: 10.13136/isr.v16i16S.970 2026, 16(16S), pp. 541-564

Data Donation From YouTube and TikTok Users: How We Implemented It

Andrea Russo[a], Dario Pizzul[b]

Abstract

The rapid evolution of the Internet continues to reshape the methodological landscape of social science research. Digital methods, although relatively new, have already undergone several key transformations. From the initial "follow the medium" paradigm, enabled by early 2000s website architectures and open APIs, to the current post-API context, marked by platform-driven restrictions on data access, researchers now face increasing constraints in obtaining platform-generated data. In response, a growing body of scholarship advocates for a shift toward user-centric approaches, with data donation emerging as a particularly promising methodological alternative. Data donation involves participants voluntarily sharing their digital trace data, offering rich, contextualized insights into online behavior while upholding ethical standards. Despite its potential, practical knowledge on how to implement data donation in empirical research remains limited.

This methodological article contributes to filling that gap by presenting the practical experience developed in the AlgoFeed project, in which 240 participants were involved in a study that included the donation of their YouTube and TikTok data to investigate feedback loops between algorithms and user behavior. We detail the project's sampling strategy, data handling procedures, enrichment techniques, and legal considerations under the European regulatory framework.

This practical implementation illustrates how data donation can help overcome platform-imposed barriers while enabling more participatory and ethically sound digital research.

Keywords: data donation, digital methods, TikTok, YouTube, computational social science.

[a] University of Pavia, Pavia, Italy.
[b] University of Pavia, Pavia, Italy.

Corresponding author:
Andrea Russo
E-mail: andrea.russo@unipv.it

Received: 21 May 2025
Accepted: 7 January 2026
Published: 12 May 2026

1. Introduction

The Internet evolves rapidly. Consequently, scholars aiming to study the Internet must continuously adapt and update their methodological approaches at a similar pace. Although digital methods have a relatively short history compared to other methodological traditions in the social sciences, they are arguably undergoing their third major evolution (Caliandro, 2024).

In the late 2000s, early digital methods works focused on widely used online platforms such as Wikipedia, Google, and user blogs, exploring how knowledge developed online (Rogers, 2019). This phase saw the introduction of the now well-known principle of "follow the medium." In the early 2010s, the widespread diffusion of social media platforms offered new avenues for inquiry—primarily through their APIs—which enabled researchers to collect large-scale datasets to study user interactions (Rieder, 2013). A third phase began in 2018 with the Cambridge Analytica scandal, which led to the progressive closure of APIs (Bruns, 2019). While framed as a move toward better data protection, this shift arguably served as a pretext for platforms to restrict access, effectively ending the era in which the medium could be easily "followed" for research purposes (Perriam et al., 2020).

To continue working with digital data, alternative strategies comprehended either collaborating with platforms or find technical workarounds to the closing—both of which entail significant limitations (Araujo et al., 2022). More recently, a growing body of literature has advocated for strengthening collaboration with a key actor that has remained in the background of digital methods research: the user. Scholars have started to emphasize the benefits of involving users more directly in the research process, leading to the proposal of a new principle for digital methods: "follow the user". This principle suggests drawing inspiration from users' digital behaviors to gather and interpret richer, more contextualized data (Breuer et al., 2023; Caliandro, 2024).

In line with this renewed focus on users, the concept of *data donation* has received increasing attention (Welbers et al., 2024; Zannettou et al., 2024). Some software tools have been developed to support this process (Araujo et al., 2022; Boeschoten et al., 2023; Pfiffner et al., 2024), yet its full potential remains underexplored (Pfiffner & Friemel, 2023). Data donation, understood as the "user-centric approach in which research participants donate their existing digital trace data to researchers" (Ohme et al., 2024, p.128), offers a promising avenue to overcome the limitations imposed by digital environments that are increasingly controlled by platforms. In Europe, recent changes in the legal framework may further support this approach by granting users greater control over their personal data.

Despite growing interest, a solid body of knowledge on data donation, particularly concerning its practical implementation, still needs to be developed. To contribute to this effort, this methodological paper presents the data donation approach adopted in the AlgoFeed project. The project involved 240 participants who donated their YouTube and TikTok data, enabling the study of "feedback loops" between users and algorithms as mediated by platform infrastructures and their socio-cultural effects.

In the following sections, we first outline the evolution of digital methods. We then introduce the concept of data donation, with particular attention to the European legal framework that supports it. Next, we explore the advantages and limitations of this approach by examining two studies that focused on YouTube and TikTok data. Finally, we present the AlgoFeed project in detail, describing its key components—sampling strategy, data donation process, data management and enrichment, and legal aspects of data protection. We conclude by reflecting on how these insights may inform future research and contribute to the broader development of data donation methodologies in the social sciences.

2. The Evolution of Digital Methods

The digital methods paradigm is grounded in the notion that the Internet should be approached less as an object of study and more as a source of methods (Rogers, 2013). This epistemological stance underpins the well-known principle of "following the medium," which encourages researchers to learn from, and repurpose, online environments, such as search engines, social media platforms, and other digital infrastructures, and their tools that inherently gather, order, classify, rank, and evaluate data. Since its introduction in the late 2000s, an evolution in three phases can be identified (Caliandro, 2021, 2024).

When first developed, this epistemological stance and its related methodological strategies largely leveraged the widespread openness and accessibility of the Internet and its data. Early research focused primarily on Google, Wikipedia, websites, and blogs, exploring how their structures influenced the development and dissemination of knowledge online (Rogers, 2019). A second phase was inaugurated around the early 2010s with the rapid diffusion of social media platforms—Facebook, Twitter, Instagram, among others. The rise of social media was accompanied by the public availability of APIs. Social media APIs (Application Programming Interfaces) are tools provided by social media platforms that allow developers and researchers to access and retrieve platform data in a structured and automated way. For research purposes, these APIs enabled scholars to collect large-scale datasets on user interactions, posts, likes, hashtags,

and network connections, facilitating the study of online behavior, communication patterns, and cultural trends (Caliandro, 2021). For example, a widely used tool at the time was Netvizz, developed in 2009 by Bernhard Rieder as an exploratory tool to study Facebook's API. Through several updates, Netvizz evolved into a comprehensive data extractor that allowed researchers to retrieve structured data from different sections of the Facebook platform in standard formats. Specifically, Netvizz provided access to three key areas: personal networks, groups, and pages. Within personal networks, it enabled the extraction of friendship graphs and "like networks," capturing user connections and affinities based on shared interests. For groups, it offered data on member interactions, creating weighted graphs based on likes and comments exchanged within group posts. In the context of pages, Netvizz generated networks linking users to posts through their engagement activities, along with tabular files that facilitated statistical and content analysis. By producing both network and tabular data files, Netvizz supported a variety of analytical approaches, ranging from Social Network Analysis to more traditional statistical methods, making it a versatile tool for studying the social dynamics and cultural patterns emerging on Facebook (Rieder, 2013).

However, the use of APIs for social science research, including Netvizz (Rieder, 2019), drastically changed after 2018, following the Cambridge Analytica scandal. The Cambridge Analytica scandal, revealed in 2018, involved the unauthorized harvesting of personal data from millions of Facebook users. The political consulting firm Cambridge Analytica collected this data—often without users' explicit consent—through a third-party app disguised as a personality quiz. This information was then used to build detailed psychological profiles aimed at influencing voter behavior in key political campaigns, including the 2016 U.S. presidential election and the Brexit referendum. The scandal sparked widespread concern over data privacy, leading to greater scrutiny of social media platforms' data practices and prompting policy changes aimed at protecting user information. These changes also affected academic research, drawing criticism and complaints from scholars (Bruns, 2019).

As a result, digital methods found themselves having to engage with a medium they intended to "follow" *but that no longer wanted to be followed*. However, it is also the researcher's task to respond to these shifts in the medium (Perriam et al., 2020). This prompted the emergence of a third phase in digital methods, which explored analytical strategies less reliant on APIs. Alongside methodological innovations in digital methods, such as web scraping and tracking (Garavaglia et al., 2023; Perriam et al., 2020), new epistemological frameworks have been proposed. In addition to the principle of "follow the medium," scholars have introduced the complementary principle of "follow the user". This approach encourages researchers to take inspiration from, and leverage, Internet

users' natively digital methods, intended as those practices, habits, techniques, and strategies through which users (both human and non-human) collect, store, manage, and organize their own digital data during their everyday online navigations (Caliandro, 2024).

As Caliandro (2024) explains, this further development in digital methods does not develop in a vacuum but builds on three sociological traditions. The approach draws first on ethnomethodology and its emphasis on examining how individuals use their everyday, taken-for-granted methods to make sense of social reality in the contexts where their activities naturally occur. Then, the idea of following the user also builds on Latour's actor-network theory (ANT). ANT invites researchers to conceive of social order not as a pre-existing structure but as something continuously produced through the associations and practices of diverse actors, which can be studied by "following the natives" and tracing the networks through which meaning and stability are achieved. Finally, a source of inspiration is multi-sited ethnography, particularly its call to "follow the things".

The relationship with ethnography is the one that is worth further exploring, also in light of some features of data donation that we will cover later. As Delli Paoli and D'Auria (2021) properly describe, doing ethnography in digital social spaces has been labelled in several ways, but different approaches to digital ethnography can be classified based on the type of content analyzed, whether it relates to digital traces detached from context or to contextualized social environments, and on the types of data involved, whether big data or small data. Following this classification, it is clear that follow the user, and more specifically data donation as it will be described in the upcoming paragraphs, falls somewhere in between. Of course, it does not deal with the corpus of big data that APIs once allowed researchers to gather, but it is perhaps not entirely correct to say that it deals with small data, because, if comprehensive research designs are proposed, such as the one described in this article, the corpus of available data can be quite large as well.

With respect to the type of content, follow the user most often deals with digital traces detached from defined and limited online spaces such as forums, communities, or Facebook groups. Since the follow the user principle, and consequently the practice of data donation, only draws partial inspiration from digital ethnography, it makes sense that the existing classifications of the latter do not fully apply. This is also because follow the user tends to go beyond the standard distinction between qualitative and quantitative methods and their associated features (Caliandro, 2024). Nevertheless, authors have advocated for further integration between data donation approaches and digital ethnography (Robinson & Cole, 2024), calling for an even larger role for users in guiding data interpretation.

With respect to the user's role more generally, within this recent evolution of digital methods, the practice of data donation is fully consistent with the idea of an expanded role for users in research, aligning with participatory research methods, where a stronger connection between researchers and participants is considered a significant advantage for the knowledge production process (Bergold & Thomas, 2012). Furthermore, "following the user," along with its associated methodological strategies, acknowledges users as collaborators in the research process and extensively involves ethical principles of doing research, distancing itself from the data-extractivist practices that mainly characterize platforms' treatment of users and their data online (Caliandro, 2018, 2024).

3. Data Donation

Data donation can be understood as a "user-centric approach in which research participants donate their existing digital trace data to researchers" (Ohme et al., 2024, p.128). More broadly, donation, conceived as a specific type of exchange, has been a topic of interest since the early contributions to the field of sociology (Mauss, 1966). It is typically defined as the transfer of something from its owner to another party without an expectation of something in return and without having any economic profit motive connected to it (Prainsack, 2019). However, reflections on the concept of the gift suggest that the act of giving often entails indirect reciprocity, as moral and social obligations are connected to both offering and receiving (Mauss, 1966; Prainsack, 2019). Specifically referring to data donation, it is important to acknowledge that this is not an issue exclusive to the field of social science. In the medical domain, it is a widely discussed topic, recognizing that data donation remains less developed and structured compared to other forms of donation, such as blood or organs. Even patients who are willing to donate their data for medical research often face practical and ethical complexities that make the process less straightforward (Krutzinna & Floridi, 2019). Within the field of social science research on online environments, a key development that has supported the initial establishment and early diffusion of data donation practices has been the recent evolution of certain elements within the legal framework governing the handling of personal digital data.

3.1 A New legal framework favoring data donation

Two recent legal developments have contributed to making data donation for research purposes more feasible, at least within Europe. First, the European

Union's 2018 General Data Protection Regulation (GDPR), under Article 15, establishes the Right to Data Portability. This provision allows individuals to download a copy of their personal data from social media platforms or any data controller, such as a supermarket. These data files, often referred to as Data Download Packages (DDPs), can subsequently be donated to researchers for academic purposes (Boeschoten et al., 2022; Ohme et al., 2024; Skatova & Goulding, 2019) Second, the recently introduced Digital Services Act (DSA) further supports data donation initiatives. According to Article 40(4) of the DSA, "upon a reasoned request from the Digital Services Coordinator of establishment, providers of very large online platforms or of very large online search engines shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers who meet the requirements in paragraph 8 of this Article, for the sole purpose of conducting research that contributes to the detection, identification, and understanding of systemic risks in the Union" (European Parliament, 2022). To be recognized as a "vetted researcher", as specified in Article 40(8), applicants must satisfy several conditions: they must be affiliated with an established research institution, remain independent from commercial interests, disclose their sources of research funding, ensure secure data management, and agree to make their research findings publicly available at no cost. Moreover, their research must address issues categorized as systemic risks by the DSA. According to Article 34, these risks include the dissemination of illegal content, potential threats to fundamental rights, disruptions to civic discourse, electoral processes, and public safety, as well as issues related to gender-based violence, the protection of public health and minors, and threats to individuals' physical and mental well-being (European Centre for Algorithmic Transparency, 2023). Despite these advancements, the practical implementation of researcher access under the DSA remains uncertain (Ohme et al., 2024).

3.2 Advantages and limitations of data donation

Two main categories of major benefits can be identified with respect to data donation: one related to the types of data that can be gathered, and the other concerning the active involvement of users. Regarding the information collected, it is important to recognize that data obtained through the donation of DDPs is often highly rich, and in some cases, even more comprehensive than data gathered via API access. It should be noted that APIs do not grant access to all available data on platforms; instead, they offer limited datasets determined unilaterally by the platforms themselves, leaving researchers with little control over the quality or quantity of the data they can access (Lomborg & Bechmann,

2014). In terms of user involvement, data donation emphasizes the active role of participants, who retain significant control over the data they choose to donate (Ohme et al., 2024). This process is further facilitated by software specifically designed for research purposes. For example, the tool Port enables research participants to request a digital copy of their personal data from social media platforms, store the DDP on their personal devices, and selectively choose the data points they wish to share with researchers (Boeschoten et al., 2023). Tools like Port also promote data minimization practices, allowing participants to donate only the data relevant to the research project while excluding unnecessary information from the dataset (Ohme et al., 2024). Such a level of participants control and precision is not feasible with methods like API-based research, scraping, or tracking. Another recently developed tool is the Data Donation Module (DDM), designed to facilitate the process of data donation. Functioning as a web-based interface, it enables researchers to structure and manage the entire donation workflow in accordance with ethical and methodological standards. The platform allows researchers to guide participants step by step—from providing consent to sharing data—while also specifying customized parameters for data preprocessing and filtering prior to transmission. Notably, the module supports the integration of structured digital trace data (e.g., in JSON or CSV formats) and accommodates the inclusion of self-reported information via an embedded questionnaire (Pfiffner et al., 2024).

These two categories of advantages—access to rich data and the active involvement of users—make data donation a promising approach for contemporary social science research on digital media. However, it is important to acknowledge certain limitations of this method, which correspond to the very domains that represent its strengths. With regard to data, obtaining large sample sizes through data donation is highly unlikely. Since this approach depends on the active participation of individuals, the scale of data collection is often limited, making it difficult for researchers to engage very large user samples. As a result, the volume of data available for analysis is typically smaller compared to the vast datasets that were once accessible through APIs, which facilitated large-scale, big data studies (Marres, 2015; Rogers, 2019). Instead, data donation aligns more closely with research designs that prioritize depth over breadth, making it particularly well-suited for qualitative or mixed-methods approaches. Additionally, a technical consideration is that data acquired through donation is often less standardized than data collected via APIs, which requires additional effort from the research team to process and analyse the information (Ohme et al., 2024; van Driel et al., 2022). Regarding user involvement, it is important to recognize that downloading DDPs is not always a straightforward process, and recruiting participants willing to donate their data can present significant challenges (Gomez Ortega et al., 2025; Ohme et al., 2024). A potential solution is

fostering even closer collaboration with research participants, providing support and guidance to help them successfully download and donate their data (Kmetty & Ne´meth, 2022; van Driel et al., 2022). Furthermore, an additional limitation concerns the fact that data donation is subject to local legal frameworks that regulate individuals' rights to access and share their personal data, such as the case of the European regulation previously discussed. Variations in how these regulations are implemented and enforced across different jurisdictions can affect the overall feasibility of data donation in various contexts (Gomez Ortega et al., 2025).

To further explore the advantages and limitations of data donation, as well as its practical implementation, we focus more extensively on two studies that relied on user-donated YouTube and TikTok data—two sources of information that align with those used in the AlgoFeed project. As previously mentioned, this methodological paper presents the AlgoFeed approach to contribute to the broader development of knowledge on how data donation can be implemented in research projects.

3.2.1 Donating YouTube data

Welbers et al. (2024) conducted two studies in which participants were asked to donate their YouTube consumption data via a web platform. In both cases, participants were instructed to request their data from Google and upload it through a newly developed web application. One study recruited participants through an online panel survey, while the other was carried out in a field lab at a large music festival. The research aimed to evaluate recruitment success, data validity, and the design of the donation application.

As anticipated, the online survey experienced a high dropout rate. Of the 9,523 participants who accessed the study link, 3,709 provided informed consent and passed the initial screening, but only 435 ultimately donated their data. The study confirmed existing concerns about response bias in data donation: donors were disproportionately younger, more highly educated, male, politically left-leaning, and scored higher on general trust measures. This suggests that, even within a survey sample—already prone to selection biases—specific characteristics further influence participants' willingness to donate their data. Additional factors, such as technical proficiency and privacy attitudes, may also play a role. The authors recommend collecting information on these variables early in the process, before excluding participants based on their willingness to donate, to allow for bias correction and gain deeper insights into donation motivations. Many participants who dropped out of the donation process cited either an unwillingness to donate or found the process too time-consuming. This

highlights potential shortcomings in the informed consent process, suggesting that some participants did not fully understand the procedure or its purpose. The authors propose placing greater emphasis on key aspects of informed consent to improve retention. Technical barriers were also a significant factor in participants dropout, particularly during the request and download of the Data Download Package (DDP).

A key finding from the field lab study was the advantage of conducting data donation research in a face-to-face environment. participants reported feeling more comfortable with the process, which was faster and more engaging compared to the online setting. The field lab also more effectively achieved the three main goals of data donation research: collecting data, involving participants in the process, and increasing their awareness of their digital footprints. Many participants reported gaining a better understanding of GDPR and its implications, suggesting that raising awareness can enhance engagement in data donation initiatives.

Regarding data validity, most participants felt their data accurately reflected their browsing, search, and YouTube history. However, approximately 25-30% of participants indicated that their data partially included activities belonging to other users, raising concerns about data quality in cases where platforms are shared (e.g., Netflix, shared household devices). This underscores the need for better mechanisms to determine data ownership and to address related ethical considerations.

The researchers developed an open-source application for data donation, which is available for future improvements. They emphasize the importance of modular design to enhance security and efficiency in various research contexts. Despite its potential, data donation faces persistent challenges, particularly technical barriers related to platform policies and public perceptions. There is a risk that research will be constrained to collecting digital traces that are easier to access—similarly to how social media studies often depend on platform APIs. Researchers are encouraged to critically assess when and how to incorporate data donation into their methodologies.

3.2.2 Donating TikTok data

Zannettou and colleagues (Zannettou et al., 2024) present the first empirical analysis of user engagement with short-form video content on TikTok, based on data collected through a data donation system involving 347 users and 9.2 million video views. Their findings show that users progressively spend more time on the platform, raising concerns about potential addictive behav-

iors, particularly among younger audiences. The analysis also highlights differences in engagement between videos from followed versus non-followed accounts and offers reflections on user attention patterns.

In addition to examining engagement metrics, the study addresses the design and challenges of data donation systems. Zannettou and colleagues implemented a data donation system called Social Media Donator (SMD), which provides participants with instructions on how to request their personal data from TikTok through the mobile application. The researchers anonymized sensitive personal information, such as phone numbers and email addresses, and allowed users to choose which data fields they were willing to share (e.g., excluding comments). The only mandatory field was the video viewing history, which included only the URLs of the videos watched and their timestamps.

Following data donation, participants were presented with an optional survey that included general demographic questions as well as inquiries about their usage of TikTok and perceptions of its recommendation algorithm. This survey offered additional contextual data on participants' age, gender, and location. Recruitment was conducted via two channels: announcements on Twitter and Facebook Ads targeting users aged 18 and older residing in the

U.S. Using these methods, the researchers recruited 347 participants, who were compensated with a total of $6,900 in Amazon gift cards delivered via email. The study also examines the role of monetary incentives in data donation, suggesting that pricing mechanisms can significantly influence individuals' willingness to donate data. Furthermore, the credibility of the data donation system proved essential for participants trust. Providing robust anonymization tools and allowing participants to control which data they shared increased user confidence in the process.

The research also highlights the need for compliance audits to verify the accuracy and completeness of data provided by social media platforms. The researchers discovered missing data in their dataset, underscoring the importance of verifying the integrity of data sources. Despite offering valuable insights, the study recognizes its limitations. The relatively small sample size and lack of comprehensive demographic data limit its ability to provide a fully representative analysis of TikTok's user base.

Additionally, the research does not account for user retention dynamics or those who stop using TikTok altogether, which would require a different data collection methodology. Some engagement metrics, such as time spent on videos, rely on inference, meaning that certain results should be interpreted as lower-bound estimates.

4. AlgoFeed Project: a practical example of data donation

As shown, data donation is receiving increasing attention in the literature as a promising solution to the challenges of conducting digital research in the post-API era (Boeschoten et al., 2023; Ohme et al., 2024; Pfiffner & Friemel, 2023). However, only a limited number of studies have provided practical insights into how data donation can be effectively implemented in empirical research (Bechmann et al., 2025; Welbers et al., 2024; Zannettou et al., 2024). We therefore aim to contribute to this discussion by presenting a practical approach to data donation, describing the activities carried out within the AlgoFeed project, in which one of the authors was actively involved in managing the data collection of participants' YouTube and TikTok data through data donation. Given the sensitive nature of the data involved, the project underwent a multistage approval process by the university's ethics committee, which carefully evaluated the protocols for informed consent, data minimization, participant anonymity, and the secure handling of personal information.

4.1 Brief summary of the project

The AlgoFeed project aims to study the impact of recommendation systems on online content consumption, deeply investigating the "feedback loops" between users and algorithms, which are mediated by the socio-technical infrastructures of platforms and their socio-cultural consequences. To achieve this, the project focuses on a quota sample of 240 adult volunteers (aged 18 to 40) who are users of YouTube and TikTok, recruited in Lombardy and Campania with the collaboration of the Demetra Opinioni agency. Our mixed-method strategy is structured into three main research phases:

A preliminary survey aimed at measuring the digital skills and consumption habits of participants, thus supervising the chosen sample so that it is as homogeneous as possible.

A data donation campaign designed to encourage participants to share data on personalized recommendations from TikTok and YouTube, in order to longitudinally and aggregately monitor their evolution and distribution within the sample.

Qualitative follow-up interviews with a subsample of 40 participants, with the goal of shed-ding light on how platform users understand and interpret algorithms and their effects on content consumption.

4.2 Sampling strategy and data collection

The sampling was conducted on a quota-based sample of 449 adult volunteers. But following questionnaires and interviews the number has been reduced to 240 adults, aged 18–40, who are active users of both YouTube and TikTok and reside in Lombardy (n = 120) and Campania (n = 120). Participants were carefully selected during the sample design phase using screening questions to verify their age and place of residence. They were identified and contacted by an external statistical agency specializing in social research (Demetra Opinioni). The external statistical agency helped us avoid potential problems that were raised in other studies (Welbers et al., 2024), specifically biases arising from convenience sampling or specific online/offline recruitment channels. Furthermore, it helped ensure that participants received an appropriate incentive (a C10 shopping voucher) to encourage active participation and data donation, making the process resemble the design of campaigns typically used in survey research based on statistical sampling procedures.

In fact, sampling quotas considered geographical distribution, gender, age, and education level.

Participation was entirely voluntary and not coerced in any way. Refusal to participate did not result in any negative consequences, whether economic or otherwise. Furthermore, participants were free to withdraw from the study at any time without providing an explanation. In the event of withdrawal, any previously collected data were destroyed, unless they had already been processed for research purposes.

The activity was divided into two phases. The first consisted of a structured questionnaire based on several sets of questions aimed at understanding participants' socio-demographic situation (e.g., educational level and gender), participants theoretical and practical digital skills, and participants browsing habits on digital platforms.

The second phase involved analysing participants' activity on YouTube and TikTok over a three-month period. Data collection for this phase was conducted through a data donation procedure. Specifically, participants were asked to donate data related to their online navigation on TikTok and YouTube, particularly concerning personalized content recommendations, to examine the influence of algorithms on online consumption.

The collected data included textual content, video content, and social media usage habits. To this end, respondents were asked to download their data from their own YouTube and TikTok profiles independently. Two simple video tutorials that had been provided to participants showed the correct procedures. The two videos (one for YouTube and one for TikTok) were created and pub-

lished on YouTube, then shared with the participants of the data donation initiative. These videos guided participants step by step through the process of requesting their data from the platform and submitting it to the AlgoFeed project. Once this process was completed, participants were asked to send the downloaded data to the research team via a dedicated email address associated with the University of Milan.

Respondents' involvement in the project lasted about four months (the total duration of the project was approximately two year), during which the respondents were required to complete a questionnaire lasting approximately 60 minutes with 37 questions, to monitor their browsing activity on YouTube and TikTok.

4.3 Donating YouTube and TikTok data

As previously mentioned, Step 2 consisted of a voluntary data donation campaign designed to encourage 240 participants to share digital data on the personalized video content recommendations they received on YouTube and TikTok in the months preceding the research. This step aimed to longitudinally and aggregately monitor the evolution and distribution of these recommendations within the sample. This research phase was coordinated by research teams from the University of Pavia, the university of Federico II of Napoli and the University of Milan, with technical support from the non-profit organization AI Forensics, which handled data collection, integration, cleaning, and anonymization.

For YouTube, participants had the option to follow the platform's official procedure (described in the video) to download specific data from their profiles, including personalized content recommendations, which were at the core of the AlgoFeed project.

For TikTok, participants followed a similar procedure, guided by a video tutorial specifically created to walk them through the simple 3–4 click process required to download the list of recommended videos from the months preceding the request.

The donated data consisted exclusively of lists of video identifiers uploaded/viewed on both platforms, meaning they did not include sensitive personal information from participants' profiles and were therefore not sufficient for identification. The right of users to obtain data stored online by platforms was guaranteed in the European Union under Article 20 of the GDPR.

The two files (one for YouTube, and the other for TikTok) provided by the platforms were then shared by participants, after signing the informed consent form, through a dedicated online form presented alongside the preliminary survey.

This process was managed by both universities, while AI Forensics enriched the dataset with metadata on algorithmically recommended content obtained via scraping and/or official APIs (e.g., video title and description, publication date, number of likes and views). Our work, in the end, consisted of:

-Anonymize the dataset.

-Management of the dataset.

-Automatically remove any data not directly necessary for research purposes, without requiring human review, to maximize privacy protection for participants.

The anonymized YouTube and TikTok datasets were then integrated with the dataset from the preliminary survey through an identification code. Only the AlgoFeed team had access to the key linking participants' identities to their data.

The resulting datasets were analysed exclusively at an aggregate level using digital and computational methods such as text mining, network analysis, and time series analysis. The objective was to descriptively map the longitudinal evolution of algorithmically recommended content types across platforms and socio-demographic categories.

The computational and digital analysis of the data from the data donation process (integrated with survey data) was conducted by the data controller in aggregate form and was limited to metadata of direct scientific interest. Any metadata obtained from YouTube and TikTok that was not necessary for research purposes was automatically and non-intrusively deleted from the final datasets

4.4 Managing and cleaning the data

Although we used the videos to instruct and guide participants through the various steps of data donation, errors emerged in both the format and type of data received. In the videos, participants were only advised not to select items related to messaging interactions with other users or economic transactions carried out on the two platforms. However, this guideline was not always followed.

As a result, during the analysis phase, we encountered five different file formats (CSV, JSON, TXT, HTML, and XLSX) while we only ask for two, and several unauthorized pieces of information, which were promptly deleted due to privacy concerns.

After collecting all the data, work began on unifying the contents into a single format (CSV) using a key-value logic structure. For each key (variable), its corresponding values (data points) were extracted from each file type. However, this process led to two main issues:

The keys were dissimilar across different file formats and platform.

Overlapping keys from different users within the same database caused conflicts, overwriting pre-existing entries.

To address these challenges, extensive data management efforts were carried out. The approach involved segmenting various IDs based on their key-value structures, followed by a final step of consolidating the data to unify individual records.

The complete dataset includes 109 respondents with a total of 3.887.463 datapoints:

3.413.299 for YouTube (videos watched, searches made, ads served), and 474.164 for TikTok (Video Like List, Favorite Video, Favorite Effect, Favorite Sound, Favorite Hashtag). Of these, the unique datapoints (URLs) are 1.097.760 for YouTube, and 415.658 for TikTok.

4.5 Enriching the data

After creating the original base dataset for both platforms, the dataset was sent to AI Forensics for further enrichment. Data enrichment refers to the process of adding additional metadata to existing content, such as YouTube or TikTok videos. Since data is collected by AI Forensics through a combination of API calls and web scraping, enrichment is only available for videos that are publicly listed and have not been removed from the platforms.

We focused on a 12-month time frame ending on the first day of data collection, spanning from July 2023 to July 2024. Table 1 shows the distribution of values across different time frames. For YouTube, we sent a dataset of 369,785 unique videos and received enrichment data for 313,627 of them. However, we included only the unique videos that were actually watched (excluding searches). For TikTok, we focused on videos from the ”Liked” and ”Favorites” lists.

Table 1: YouTube and TikTok Video Counts over Different Timeframes

Timeframe	YouTube	TikTok	Total
2 years	1,744,784	389,866	2,134,650
1 year	1,126,554	311,490	1,438,044
6 months	643,890	221,060	864,950
3 months	368,590	144,405	512,995

For YouTube more specifically, we sent a list of video IDs, and received a dataset including the following metadata for each video as shown in Table 2.

Table 2: Distribution of Videos Across Respondents

Metric	Complete Dataset	12-Month Dataset
Unique IDs	104	98
Total Videos	2,336,357	682,956
Enriched Videos	1,068,257	641,366
Overall Percentage	45.72%	93.91%

The enrichment keys for YouTube are shown in Table 3.

Table 3: YouTube Dataset Field Descriptions

5.1. Field	5.1. Description
5.1. video id	5.1. Unique identifier of the video
5.1. reference	5.1. Reference associated with the video
5.1. kind	5.1. Type of resource
5.1. etag	5.1. ETag identifier of the resource
5.1. published at	5.1. Video publication date
5.1. channel id	5.1. YouTube channel identifier
5.1. channel name	5.1. YouTube channel name
5.1. title	5.1. Video title
5.1. description	5.1. Video description
5.1. category id	5.1. Video category
5.1. live broadcast content	5.1. Live broadcast status
5.1. default language	5.1. Default language of the video
5.1. default audio language	5.1. Default audio language
5.1. country	5.1. Country of origin of the video
5.1. tags	5.1. Tags associated with the video
5.1. view count	5.1. Number of views
5.1. like count	5.1. Number of likes
5.1. favorite count	5.1. Number of times the video was added to favorites
5.1. comment count	5.1. Number of comments

5.1. thumbnails	5.1. URL of the video thumbnails
5.1. channel thumbnails	5.1. URL of the channel thumbnails
5.1. duration	5.1. Video duration
5.1. dimension	5.1. Video dimension (2D/3D)
5.1. definition	5.1. Video quality (SD/HD)
5.1. caption	5.1. Presence of subtitles
5.1. licensed content	5.1. Indicates whether the content is licensed
5.1. content rating	5.1. Content rating
5.1. projection	5.1. Type of projection (normal/360°)

For TikTok, we sent 272,409 unique TikTok Video IDs, and received 230,211 unique videos enriched with metadata show in Table 4.

Table 4: TikTok Dataset Field Descriptions

2.2.1. Field	2.2.2. Description
2.2.3. display name	2.2.4. User's display name
2.2.5. country code	2.2.6. User's country code
2.2.7. video description	2.2.8. Video description
2.2.9. music id	2.2.10. ID of the associated music
2.2.11. like count	2.2.12. Number of likes received
2.2.13. comment count	2.2.14. Number of comments received
2.2.15. share count	2.2.16. Number of shares
2.2.17. view count	2.2.18. Number of views
2.2.19. effect ids	2.2.20. IDs of applied effects
2.2.21. hashtag names	2.2.22. Hashtags associated with the video
2.2.23. playlist id	2.2.24. Playlist ID
2.2.25. voice to text	2.2.26. Text generated from voice transcription
2.2.27. id	2.2.28. Unique video identifier
2.2.29. create date	2.2.30. Video creation date
2.2.31. duration type	2.2.32. Type of video duration
2.2.33. favorites count	2.2.34. Number of times the video was favorited
2.2.35. stem verified	2.2.36. Audio track verification status

5. Personal Data and GDPR

Significant attention was dedicated to data protection throughout the entire research process, with efforts made to ensure that participants were properly informed. All information regarding the processing of participants' personal data—including special categories of data—was provided in a dedicated information notice prepared pursuant to Article 13 of Regulation (EU) 2016/679 (GDPR).

All data were collected and analysed exclusively for academic and public dissemination purposes and were not shared with third parties. The legal basis for processing was the explicit consent of the data subject, in accordance with Article 6(1)(a) of Regulation (EU) 2016/679. Respondents could withdraw their consent at any time by contacting the University or the scientific lead of the

project. In such cases, continued participation in the research was no longer possible.

The survey was conducted in compliance with ethical guidelines for data collection in social research and with the GDPR regulations on data protection and privacy. The data files did not contain any personally identifiable information, but only anonymized identifiers, which allowed for the linkage of datasets between the first and second phases of the research. The data controller securely stored the data in dedicated databases or on encrypted drives, ensuring that data management was fully compliant with GDPR requirements. At the conclusion of the research project, the data controller will ensure that all databases are destroyed within two years. Access to the data is granted solely for the purpose of reviewing the publications resulting from the project.

6. Conclusion

Data donation is receiving growing attention in the literature as a promising solution for addressing the challenges of digital research in the post-API era (Boeschoten et al., 2023; Ohme et al., 2024; Pfiffner & Friemel, 2023). However, only a limited number of contributions have shared practical insights on how to implement data donation within empirical research (Welbers et al., 2024; Zannettou et al., 2024). This paper has aimed to add a valuable example to this emerging field by illustrating how data donation was implemented in the AlgoFeed project to collect YouTube and TikTok data. We have described key aspects of the process, including the use of monetary incentives, the creation of video tutorials to support participants in retrieving their data, and the data management procedures adopted—along with the challenges we encountered. We have also shared the data enrichment work conducted by AI Forensic, which significantly enhanced the project's analytical potential. Furthermore, we have outlined the European legal framework that supports this type of user-centric data collection, which may be practically useful for researchers undertaking similar work. Ultimately, we have shown an example of the valuable insights that this type of data can provide.

While the AlgoFeed experience has further demonstrated several advantages of data donation, such as greater participant involvement and access to granular digital trace data that would otherwise remain inaccessible, it has also highlighted some important limitations. Data management was not straightforward, especially for researchers accustomed to API-based workflows. Furthermore, despite video tutorials, many participants encountered difficulties in retrieving their data. This challenge emerged even within a relatively young

sample, which should be - on average - digitally literate, therefore raising concerns about the feasibility of similar approaches with less digitally proficient populations. Moreover, a substantial proportion of participants either did not complete the donation process or provided incomplete data (only 109 out of 240 provided usable data profiles), a challenge also reported in other studies employing data donation approaches (Welbers et al., 2024).

These limitations, along with the overall experience of the AlgoFeed project, suggest several reflections for further developing data donation approaches. The aim is to better fulfill their potential for collecting rich digital data while ensuring a more participatory role for research participants, one that also enhances their understanding of how their personal data is used online (Welbers et al., 2024). First, it is crucial to engage participants closely—both by offering technical support during the donation process and by clearly communicating the scientific value of their contribution. With respect to technical support, other studies have developed platforms for supporting data donation (Boeschoten et al., 2023; Welbers et al., 2024), we decided to use video tutorials. These initiatives can certainly help, but having the opportunity to more directly support participants, even with the technical aspects of the donation process, would undoubtedly be beneficial. Of course, aiming for such close contact with participants would necessarily reduce the number of individuals involved. However, this trade-off can lead to improved data quality. At present, it seems preferable to obtain complete datasets from a smaller number of participants who are fully engaged in the donation process, understand its rationale and purpose, and receive adequate support in navigating its technical aspects.

Such a more limited but richer dataset can then be further enhanced through complementary approaches. The AlgoFeed project provided a valuable example of how donated data can be enriched. The enrichment activity carried out by AI Forensics appears to be a recommended strategy for strengthening the informative power of the dataset to be analysed. This was achieved by combining the personal data donated by participants with additional data obtained through other digital methods approaches, such as API calls and web scraping, effectively merging the strengths of multiple techniques.

More specifically, as shown in our contribution, when dealing with data donation, researchers should be prepared to work with unstructured or incomplete data, which may require additional effort in cleaning and preparation. Once again, these challenges can also be mitigated if participants are well supported throughout the data retrieval process and provided with clear, accessible instructions. Therefore, based on the experience of the AlgoFeed project and previous research contributions, it seems that strengthening the bond between researchers and participants is key to further improving the potential of data donation. In this respect, it is important to remember that data donation is not

just about the data itself, but also about participant involvement and awareness. This perspective aligns with a more participatory and ethical research paradigm that moves beyond extractivist models that treat participants solely as data sources (Caliandro, 2021).

7. Funding

This project was financed by the Ministry of University and Research (PRIN 2022 call (Prot. 2022YRJ83A), project name: "Feedback culture: assessing the effects of algorithmic recommendations on platformised consumption – ALGOFEED").

References

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Welbers, K. (2022). Osd2f: An open-source data donation framework. *Computational Communication Research*, *4*(2), 372–387. https://doi.org/10.5117/CCr2022.2.001.ArAU

Bechmann, A., Brems, M. K., Olesen, M. K., Walter, J. G., & Wegmann, D. (2025) Data donation as a method for investigating trends and challenges in digital media landscapes at national scale: The Danish population's use of YouTube as an illustrative case. Nordicom, University of Gothenburg. https://doi.org/10.48335/9789189864184

Bergold, J., & Thomas, S. (2012). Participatory research methods : a methodological approach in motion. *Historical Social Research*, *37*(4), 191–222. https://doi.org/https://doi.org/10.12759/hsr.37.2012.4.191-222

Boeschoten, L., Mo¨ller, J.E., Araujo, T. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, *2*, 388–423. https://doi.org/10.5117/CCR2022.2.002.BOES

Boeschoten, L., Schipper, N.C.D., Struminskaya, B., Mendrik, A.M. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, *8*, 1–8. https://doi.org/ 10.21105/joss.05596

Breuer, J., Kmetty, Z., Haim, M., Stier, S. (2023). User-centric approaches for collecting facebook data in the 'post-api age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, *26*(14), 2649–2668. https://doi.org/10.1080/1369118X.2022.2097015

Bruns, A. (2019). After the 'apicalypse': Social media platforms and their fight against critical scholarly research. *Information Communication and Society*, *22*(11), 1544– 1566. https://doi.org/10.1080/1369118X.2019.1637447

Caliandro, A. (2018). Digital methods for ethnography: Analytical concepts for ethnographers exploring social media environments. *Journal of Contemporary Ethnography*, *47*(5), 551–578. https://doi.org/10.1177/0891241617702960

Caliandro, A. (2021). Repurposing digital methods in a post-api research environment: Methodological and ethical implications. *Italian Sociological Review*, *11*(4S), 225–242. https://doi.org/10.13136/isr.v11i4S.433

Caliandro, A. (2024). Follow the user: Taking advantage of internet users as methodological resources. *Convergence*, *0*(0), 1–24. https://doi.org/10.1177/13548565241307569

Delli Paoli, A., & D'Auria, V. (2021). Digital Ethnography: A Systematic Literature Review. *Italian Sociological Review*, *11*(4S), 243–267. https://doi.org/10.13136/isr.v11i4S.434

European Centre for Algorithmic Transparency (2023). *Dsa data access for researchers european commission algorithmic-transparency.ec.europa.eu/*.

European Parliament (2022). *Digital services act.* Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/legalcontent/EN/TXT/HTML/?uri=CELEX:32022R2065#d1e40-1-1

Garavaglia, E., Caliandro, A., Melis, G., Sala, E., Zaccaria, D. (2023). Contrasting ageism in research on older adults and digital technologies. *Digital Ageism*, 248–265. https:// doi.org/10.4324/9781003323686-14

Gomez Ortega, A., Bourgeois, J., Hutiri, W.T., Kortuem, G. (2025). Beyond data transactions: a framework for meaningfully informed data donation. *AI & Soc, 40,* 1–18. *https://doi.org/10.1007/s00146-023-01755-5*

Kmetty, Z., & Németh, R. (2022). Which is your favorite music genre? a validity comparison of facebook data and survey data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 154*(1), 82-104. doi/10.1177/07591063211061754

Krutzinna, J., & Floridi, L. (2019). *The ethics of medical data donation.* Cham: Springer Open.

Lomborg, S., & Bechmann, A. (2014). Using APIs for Data Collection on Social Media. *The Information Society*, *30*(4), 256–265. https://doi.org/10.1080/01972243.2014.915276

Marres, N. (2015). Why Map Issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values*, *40*(5), 655-686. doi/10.1177/0162243915574602

Mauss, M. (1966). *The gift.* London: Cohen & West LTD.

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B.B., Robinson, T.N. (2024). Digital trace data collection for social media effects research: Apis, data donation, and (screen) tracking. *Communication Methods and Measures*, *18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

Perriam, J., Birkbak, A., Freeman, A. (2020). Digital methods in a post-api environment. *International Journal of Social Research Methodology*, *23*(3), 277–290. doi.org/10.1080/13645579.2019.1682840

Pfiffner, N., & Friemel, T.N. (2023). Leveraging data donations for communication research: Exploring drivers behind the willingness to donate. *Communication Methods and Measures*, *17*(3), 227–249. doi.org/10.1080/19312458.2023.2176474

Pfiffner, N., Witlox, P., Friemel, T.N. (2024). Data donation module: A web application for collecting and enriching data donations. *Computational Communication Research*, *6*(2), 1, 1-24. https://doi.org/10.5117/CCR2024.2.4.PFIF

Prainsack, B. (2019). Data Donation: How to Resist the iLeviathan. In J. Krutzinna (Eds.) et. al., *The Ethics of Medical Data Donation*. (pp. 9–22). Cham: Springer. https://doi.org/10.1007/978-3-030-04363-6_2

Rieder, B. (2013). Studying facebook via data extraction: The netvizz application. *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, 346–355. https://doi.org/10.1145/2464464.2464475

Rieder, B. (2019). Netvizz is no longer publicly accessible and will very probably not come back in the future. *The Politics of Systems*, http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independentresearch-just-got-a-lot-harder/

Robinson, J. Y., & Cole, S. (2024). *Proposing reciprocal digital methods: A user-centric method for algorithmic social media platforms in a post-API world* [Paper presentation]. AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. http://spir.aoir.org

Rogers, R. (2013). *Digital methods*. Cambridge, MA: MIT Press.

Rogers, R. (2019). *Doing digital methods*. London, UK: Sage.

Skatova, A., & Goulding, J. (2019). Psychology of personal data donation. *PLoS ONE*, *14*(11), 1–20. https://doi.org/10.1371/journal.pone.0224240

van Driel, I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures*, *16*(4), 266-282. https://doi.org/10.1080/19312458.2022.2109608

Welbers, K., Loecherbach, F., Lin, Z., Trilling, D. (2024, January). Anything you would like to share: Evaluating a data donation application in a survey and

field study. *Computational Communication Research*, *6*(2), 1-25. https://doi.org/10.5117/CCR2024.2.5.WELB

Zannettou, S., Nemes-Nemeth, O., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., & Roesner, F. (2024). Analyzing user engagement with tiktok's short format video recommendations using data donations. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–16. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642433