OSD2F: An Open-Source Data Donation Framework

Araujo, T.; Ausloos, J.; van Atteveldt, W.; Loecherbach, F.; Moeller, J.; Ohme, J.; Trilling, D.; van de Velde, B.; de Vreese, C.; Welbers, K.

DOI
10.31235/osf.io/xjk6t
10.5117/CCr2022.2.001.ArAU

Publication date
2022

Document Version
Final published version

Published in
Computational Communication Research

License
CC BY

Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.2 (2022) 372-387
HTTPS://DOI.ORG/10.5117/CCR2022.2.001.ARAU

OSD2F: An Open-Source Data Donation Framework

Theo Araujo
Amsterdam School of Communication Research (*ASCoR*), *Department of Communication Science, University of Amsterdam*

Jef Ausloos
Institute for Information Law, University of Amsterdam

Wouter van Atteveldt
Department of Communication Science, Vrije Universiteit Amsterdam

Felicia Loecherbach
Department of Communication Science, Vrije Universiteit Amsterdam

Judith Moeller
Amsterdam School of Communication Research (*ASCoR*), *Department of Communication Science, University of Amsterdam*

Jakob Ohme
Amsterdam School of Communication Research (*ASCoR*), *Department of Communication Science, University of Amsterdam*

Damian Trilling
Amsterdam School of Communication Research (*ASCoR*), *Department of Communication Science, University of Amsterdam*

Bob van de Velde
Independent Researcher

Claes de Vreese
Amsterdam School of Communication Research (*ASCoR*), *Department of Communication Science, University of Amsterdam*

Kasper Welbers
Department of Communication Science, Vrije Universiteit Amsterdam

Abstract

The digital traces that people leave through their use of various online platforms provide tremendous opportunities for studying human behavior. However, the collection of these data is hampered by legal, ethical, and technical challenges. We present a framework and tool for collecting these data through a *data donation platform* where consenting participants can securely submit their digital traces. This approach leverages recent developments in data rights that have given people more control over their own data, such as legislation that now mandates companies to make digital trace data available on request in a machine-readable format. By transparently requesting access to specific parts of this data for clearly communicated academic purposes, the data ownership and privacy of participants is respected, and researchers are less dependent on commercial organizations that store this data in proprietary archives. In this paper we outline the general design principles, the current state of the tool, and future development goals.

Keywords: digital trace data, data donation, digital platforms, privacy

OSD2F: An Open-Source Data Donation Framework

Computational social science promises that digital trace data can be used to study human behavior and interaction at an unprecedented level of detail (Lazer et al., 2009, 2021). These data are generated while using digital platforms – such as Google, Apple, Facebook, Microsoft, Amazon, among many others – and encompass an ever-growing set of life domains – for example interpersonal communication, politics, commerce, health, or work.

While these data can be crucial to answer critical social scientific research questions, access is a major challenge: These digital traces are closed off in proprietary archives of commercial corporations, requiring researchers to *work with* the platforms – using their APIs or data-sharing initiatives – or *work around* them – scraping data directly, or partnering with users (Ausloos & Veale, 2020; Bruns, 2019; Freelon, 2018; Halavais, 2019).

Partnering with users, in particular, has gained momentum due to the expansion of users' access and portability rights (mandated, for example, by the EU's General Data Protection Regulation; GDPR). This turns the

researcher-platform-user relationship around: Instead of relying on platforms to study users (often without their knowledge), researchers work directly with users willing to donate their data to academic research (for an overview, see Ausloos & Veale, 2020; Boeschoten et al., 2020). This enables meaningful informed consent and linking these digital traces to self-reports (e.g., survey answers). More broadly, it cultivates awareness about digital infrastructures, fostering a culture of citizen science.

We present the *Open-Source Data Donation Framework* (OSD2F), a reusable, adaptable, general-purpose tool to facilitate the data donation process in a transparent, privacy-protecting, and flexible manner. It has the potential to foster a researcher-user collaboration to collect crucial data to study causes, contents, and consequences of (online) communication within platforms.

Accessing Digital Trace Data for Academic Research: Related Approaches

Researchers have adopted several methods to partner with users to collect and make sense of digital trace data for academic research. Often, this has been done as a complement or alternative to working *with* platforms, given restrictions and lack of transparency (Bruns, 2019; Halavais, 2019). Broadly speaking, two main data collection modes have been deployed: collecting real-time data or the donation of existing digital trace data.

Collecting real-time data

One way to collect digital trace data is to have users install specific software, plugin or proxy that monitors the user behavior and collects real-time data (for an overview, see: Christner et al., 2021). This approach tracks individual respondents' online behavior – such as pages visited in a browser, or apps used on a smartphone – and has often been used in research. Panel participants either have the software already installed (e.g., Araujo et al., 2017; Scharkow, 2016) or install it for the purposes of the study (Verbeij et al., 2021). Other studies developed their own software or browser plugins (e.g., Bodó et al., 2017; Kobayashi & Boase, 2012; Reeves et al., 2020), retaining a higher level of control in the data collection[1]. In some cases, this software may even *generate* data to study platforms, as in the AlgorithmWatch's project tracking Google results on a regular basis (Puschmann, 2019).

While this approach provides interesting options for research with digital trace data, it is restricted by the need to install software (and the device- and version dependence of such efforts) and lacks access to retrospectively capture user data. In addition, these approaches may require additional levels of privacy and data protection measures for the collected data, given

the risk of spillover of third-party information (for an overview of this and related issues, see Lazer et al., 2021) and the sheer volume of data collected about an individual, especially in instances where participants may only provide a blanket consent (i.e., for all data of a specific type to be collected, without participant review of each individual item), or in which the real-time tracking method may in some instances collect information beyond what is strictly needed for the study (e.g., capturing regular screenshots of a mobile phone or full URLs of all pages visited by a participant).

Donation of existing data

Another promising approach is the donation of *existing* digital trace data. Studies focusing on mobile devices have relied on respondents manually taking screenshots of the Screen Time function (Ohme et al., 2020), battery usage per app (Baumgartner et al., 2021), or downloading their text message history (Brinberg et al., 2021). Other studies used browser plugins such as Web historian (Menchen-Trevino, 2016), that allow users to visualize and donate their browsing history data.

The growing awareness of "data rights" – including most importantly the rights for individuals to access and transfer the data that digital platforms or companies have on them – opens several opportunities for academic research with this approach (Ausloos & Veale, 2020). These rights – established in the EU's GDPR among other jurisdictions – mandate platforms or companies to make user data available in a machine-readable format for individuals that request it. Such Data Download Packages (DDPs) may include usage information (e.g., login history), activity (e.g., posts created, messages sent) and profiling done by the platform (e.g., inferred interests or categories), among other information[2].

This approach offers several advantages. First, it relies on an explicit consent between researcher and user and increases transparency about the data shared, as users have the possibility to review their data before the donation. Second, relying on DDPs provided by platforms avoids the installation of additional software. This saves development resources and might include a broader user base, including less tech-savvy users. Third, building on user's data rights, these donations provide access to digital trace data retroactively (e.g., one's earlier activities) and circumvent concerns about reactivity (e.g., knowledge that a tracker is installed altering one's behavior). OSD2F is built to make use of these advantages, allowing researchers to work with individuals to download and donate their digital trace data by exercising their data rights, as outlined below[3].

The Framework

Intended Purpose

OSD2F enables researchers to collect and analyze individual-level digital trace data by working with users who are willing to download and donate their data for academic research. It provides a web-based interface where participants can (a) get instructions on how to download their data from specific platforms, (b) visualize the contents of their DDP, (c) (de)select the items that they (do not) wish to share and (d) donate the data to the specific research project. In the background, OSD2F follows a set of pre-processing steps to minimize the data before it reaches the researchers. In addition, OSD2F can be integrated in a survey flow, enabling researchers to combine self-reports (answers to a survey) with digital trace data (donated via OSD2F) in the same project. The overall process is outlined in Figure 1.

Design principles

OSD2F places the individual as a data subject, maintaining agency over their data. As such, it aims to enable data access by researchers in a way that respects individual rights. It is built with privacy, transparency, and flexibility as key design principles, as outlined below.

Data minimization

OSD2F is designed to constrain access to the data that is necessary for the purposes of the *legitimate interest* of researchers, i.e., required for the academic study in question, and takes steps to safeguard the privacy and data protection of participants, through four data minimization levels:

1. Researchers have to specify which files from a given DDP are needed for the research (e.g., posts, comments, login history) in an allowlist. For participants, this means that, when they load the DDP in the web-based data donation interface, OSD2F runs local scripts to read the DDP and extract only files specified by the researcher, which are then sent for further processing on the server. Other information in the DDP is ignored.
2. Researchers have to specify which fields in DDP files are of interest to the research (e.g., the text of a post) and which scripts should be executed to anonymize the data (e.g., removal of names by string matching). For participants, this means that any field not on the allowlist is removed from the files that they load on OSD2F (locally) before donated data is sent to the researcher-controlled server.

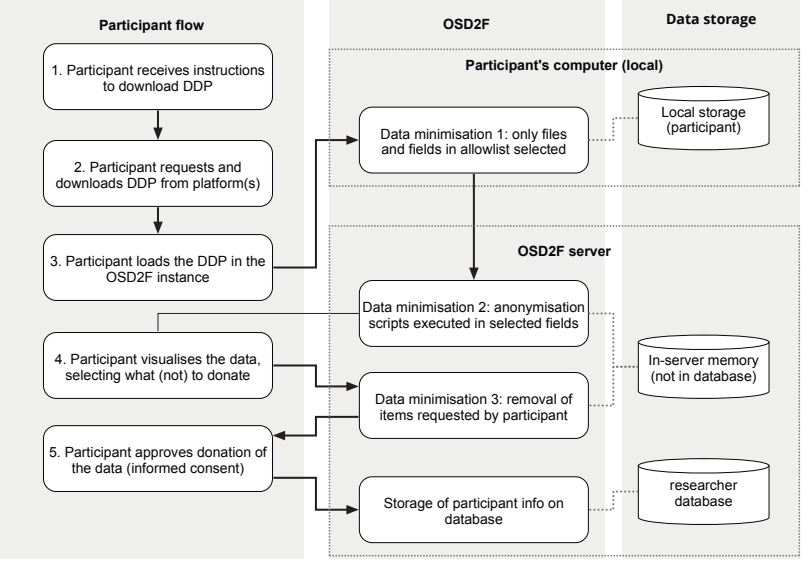

Figure 1. Overview of OSD2F flow

3. When the DDP uploaded by the participant reaches the server, each file is then processed by a set of customizable scripts that anonymize, when applicable, its contents, or perform additional analysis steps (e.g., extract only specific information or apply a classifier to categorize the information).
4. Finally, participants visualize online the subset of data from their DDP that could be donated to the academic study. They can remove items that they do not wish to donate prior to providing informed consent, with OSD2F only recording aggregate-level information (e.g., number of items removed).

Transparency

Transparency on what is being donated is central to OSD2F's design principles, as is ensuring individual agency during the process. Drawing from the design principles proposed by, for example, Web Historian (Menchen-Trevino, 2016), participants can *visualize* which subset of their data might be donated to the academic study. In this process, they can remove specific items (see Figure 2). For those, OSD2F only stores aggregate information (e.g., number of posts removed). After the review, the participant is asked to provide informed consent, approving the donation of the data. *After* explicit

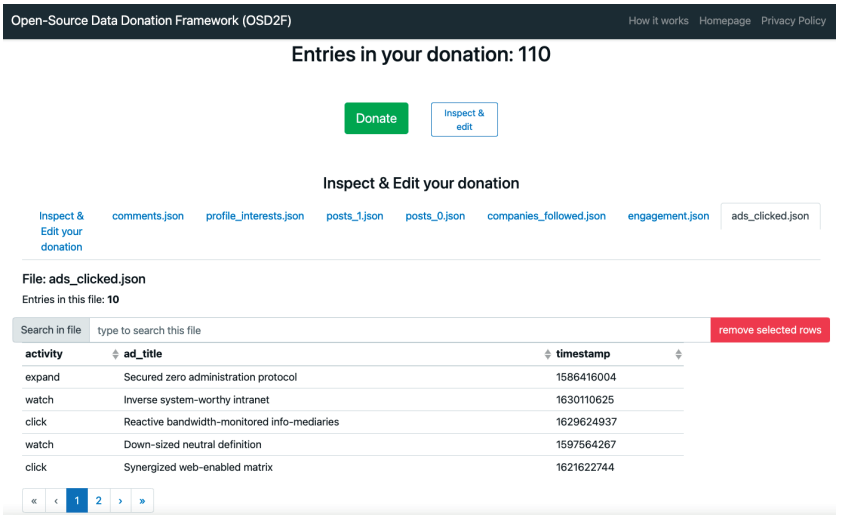

Figure 2. Example of the DDP inspection page

approval by the participant, the data – that up until that moment were only kept in memory of the OSD2F instance – are saved in the database.

Flexibility

OSD2F is designed with flexibility as a key principle and developed in Python – given the popularity of the language within the academic community – and enables customization *both* of its user interface (UI) and of how it handles DDPs. Importantly, it is designed to be installed and configured *per study* by a researcher. This means that the instructions provided to the participant and how the DDP is handled are specifically tailored to the study at hand.

All pages (e.g., privacy policy, data donation page, DDP inspection page etc.) and elements in the UI are generated based on a configuration file (see Figure 3) to allow for the configuration of the whole participant-facing interface. This also allows for the translation to the language appropriate for the study.

The DDP handling is also managed through a configuration file (Figure 4) where the researcher specifies the filenames to be extracted from the DDP and, per filename, the fields of interest with the corresponding anonymization script to be executed per filename (if required). This is used for the data minimization steps (described above), and works with the principle of *allowlisting*, i.e., only items (filenames, fields) explicitly specified are collected. The flexibility in handling DDP formats not only ensures that new platforms with JSON-format DDPs can be added, but also allows researchers to react quickly to the frequent changes in DDP formats by platforms.

```
project_title : OSD2F
contact_us : email@domain.tld
static_pages:
  home:
    active : True
    name : The OSD2F project
    blocks:
      - type : jumbotron
        title : Data donation made easy
        id : top
        image : "/static/skull_phone_cc.jpg"
        lines :
            - A general way of donating data
            - For JSON based GDPR exports
            - To use with external survey and analysis tools
        buttons :
            - name : About the project
              label : "btn-primary"
              link : "#project"
            - name: How it works
              label : "btn-success"
              link : "/donate"

      - type : two_block_row
        id : project
        image: "/static/study_cc.jpg"
        image_pos: left
        title : OSD2F provides a whitelist based collection website
        lines :
          - Under GDPR, everyone should be able to export <br> their data in machine-readable format
          - Many platform provide standardized ways to get this data by <br> exporting it as a set of JSON files
          - This app allows researchers to easily and safely collect exported data donated by participants in their studies
```

Figure 3. Example of the content configuration file

Using and Extending OSD2F

General Overview

OSD2F generates a web interface containing both the interface for participants to donate their data and – if authentication is configured – also for researchers to download the donated data. Its core is written in Python – with the web server based on Flask[4]. JavaScript is used for the UI interactive elements, including (a) the steps where participants load their DDP locally, extract the allowlisted files from the DDP and upload them to the server and (b) the interface for data visualization and selection. Configuration files use the YAML format.

Technical Requirements and Installation

Detailed information about the latest requirements and installation instructions are available at the OSD2F GitHub repository[5]. Below, we describe its main components.

Server

The framework can be installed locally on the researcher's computer for development and testing purposes yet is ultimately designed to run in a web server during production, i.e., data collection with participants. The production version of OSD2F can be installed in a traditional server or in a cloud-based web app (e.g., Azure Web App, Amazon Elastic Beanstalk, Google App Engine, Heroku etc.).

```
# Sample Platform upload settings
files:
    (^|/|\\)comments.json:
      # in key is required when the initial level of data
      # is not a list but an object (e.g. {} instead of [])
      in_key : "comment_information"
      anonymizers:
        - redact_text : ""
      # accepted_fields to include in the upload
      # remove a field to filter it out
      accepted_fields:
        - timestamp
        - title
        - information.comment.comment_text
```

Figure 4. Example of the DPD handling configuration file for comments

Database

The donated data are stored in a database after participants provide informed consent. Logs with participant and researcher activity on the OSD2F instance are also stored. For development and testing, data storage can be done in a non-persistent (in-memory) SQLite database. For production, the researcher needs to set up a persistent database (any database supported by Tortoise[6]), configuring and securing access appropriately.

Deploying OSD2F

The deployment of OSD2F in general terms include two phases: development and testing, and production. We sketch below the key activities in each of these phases.

Development and Testing Phase

OSD2F comes with a set of baseline scripts to be adapted to specific platform JSON-based DDP formats of interest (for the latest details, see its GitHub repository). In the development phase, the researcher should familiarize herself with the DDP format of the platform(s) and update the *upload settings* (Figure 4, above) to (a) specify which files and fields to include in the allowlist and (b) which anonymizer script (if applicable) to execute for each file. In addition, the researcher should then update or create the appropriate anonymizer scripts (in Python). The researcher can also configure the participant-facing interface, by updating the *content settings* (Figure 3, above).

Running OSD2F locally with an SQLite database, the researcher can iteratively check if the configurations work as expected. If so, it is recommended to test the OSD2F configuration thoroughly with more examples of DDPs.

Production Phase
After the development and testing are complete, the researcher can set OSD2F for data collection with participants. At this stage, a database should be set up to store the donated data and the logs, and OSD2F should be deployed to the production server (or web app). The researcher should setup and validate with their institution all security measures (details below), the authentication for access to the researcher's page where the donated data can be downloaded[7], and the encryption key for the data stored in the database (if applicable). With this complete, the researcher can direct participants to donate their data using OSD2F.

Linkage with survey data
If using an online survey tool (e.g., Qualtrics) as part of the recruitment process, a special unique identifier per participant can be generated (e.g., a random number) and sent to OSD2F via the URL provided to the participant[8]. OSD2F will then record the donated data along with this unique identifier, allowing the researcher to later connect the survey answers to the donated data.

Considerations

The steps above provide a technical overview of how a researcher can set up an OSD2F instance. Below we briefly outline some important considerations in this process.[9]

Legal Considerations
OSD2F capitalizes on the legal requirement for online platforms granting users access to their personal data. In the EU, this requirement can be found in Articles 15 (Right of access by the data subject) and 20 (Right to data portability) of the GDPR, with similar rules found in other jurisdictions[10]. Such provisions enable donation-based research, but also raise requirements on researchers collecting this information. Before using OSD2F, researchers must therefore consider the relevant legislation applicable to them, consult with the appropriate instances in their institution and run, if needed, a Data Protection Impact Assessment (DPIA) – also for the considerations below.

Ethical Considerations
OSD2F is built around the notion of researchers partnering with users. Researchers should carefully discuss the project with their Ethical Review Board

(ERB). Such discussions should cover what will be collected (i.e., what is of *legitimate interest* for the research project), how the data will be processed and stored, and the process of participant recruitment and appropriate incentives to participate. Importantly, the informed consent form and privacy policy needs to clearly state what is collected, and for what purpose[11]. In addition, researchers should be clear on how participants can withdraw their consent and whether the data will be reused in the future (and under which conditions).

Privacy Considerations

The deployment of OSD2F raises three important privacy considerations. First, data included in most DDPs can be potentially used to *identify* participants even after some anonymization steps are taken (e.g., the content of a post, even without a username, can in some cases be used to identify the participant). Second, some DDPs contain information about *third parties* (e.g., names of friends) which may lead to information spillover (for an overview, see Lazer et al., 2021), thus the researcher should consider carefully with their ERB the extent to which an individual's informed consent can reasonably cover this information, with the alternative being an anonymizer script to remove this before storage in the database. Third, some data included in the DDP may be inadvertently *sensitive* in nature despite data minimization steps being executed (e.g., even if fields on political affiliation, sexual orientation or health status are not included in the allowlist, this could be disclosed in or inferred from the posts by the participant). Researchers should therefore agree on appropriate measures for these aspects in their discussion with the ERB and in the DPIA process, and carefully consider the extent to which the (unprocessed) data collected can be shared outside of the context of the study, especially given the potential for deanonymization. In addition, researchers should carefully assess the extent to which data collected with this approach may be *shared* across research teams or with the broader research community, and under which circumstances. While, for instance, there is widespread consensus over the benefits of large survey panel datasets available to any interested researcher (e.g., the European Social Survey), the associated risks with donated data often (but not always) are higher. They may be more prone to de-anonymization, or the linkage of different sources might result in disclosure of attributes that were not meant to be disclosed. Hence, there generally is a tension between the desire to share data (both for transparency reasons and to create synergies) and the need to protect the donators. And, in all instances, one must first understand the extent to which potential participants may provide informed consent for this.

Security and Data Management

Given the potentially private and sensitive nature of the data, researchers interested in using OSD2F are advised to carefully consider with their institutions how to secure the infrastructure during the data collection process (e.g., where to host OSD2F or the database, how to secure the database etc.) and after the data collection is complete (e.g., where the data will be stored, who will have access and under which conditions). It is advised to delete the OSD2F instance – and its database – as soon as the data collection is complete. OSD2F also provides a set of baseline security measures (e.g., ability to use a key-vault to store keys, or to store the donated data in an encrypted form, etc.), but the researcher should validate all security measures with the appropriate instances in their institution[12].

Discussion: Limitations and Future Work

OSD2F creates a series of opportunities for researchers to work with individuals who are willing to make use of their data rights and contribute to academic research. Some limitations, however, need to be considered.

First, from a technical perspective, OSD2F is designed to handle individual-level DDPs (i.e., the data one can download about oneself from a digital platform) in JSON format. It works with certain assumptions (e.g., patterns in filenames determine the type of data being donated) that on the one hand ensure flexibility (e.g., allow a researcher to configure which file to read, and which fields to use), but on the other hand may constrain the types of file or platforms being processed. We plan to expand the capabilities and flexibility of OSD2F in future releases.

In addition, the current pre-processing scripts handle each entry in each file sequentially. This means that researchers with use cases where the assumption of independent entries does not hold (e.g., exploring linkages between users on the platform, or how posts are shared across communities) would require extensions of the current scripts. This highlights one of the advantages of the proposed approach, as researchers can setup their own servers using the base OSD2F code, while developing their own preprocessing pipelines – including developing Python scripts that can anonymize or preprocess the ingested data in ways that help their specific use case, and, hopefully, sharing their results and experiences with the wider community. In addition, as researchers can configure what they extract from the DDP, they may – after required legal and ethical considerations –, for example,

decide to retrieve specific individual-level information from the DDP (e.g., URLs shared by a participant) that can be enriched in later steps of the researcher's analysis pipeline (e.g., retrieving the content of the URLs to understand what type of information they share).

Second, from the perspective of participants, the process to request a DDP from a platform may be cumbersome to those with lower technical comfort. Important aspects to consider here are the time it takes for the DDP to be delivered to the user and the information required to access the data. Online platforms can take up to 30 days to answer data requests – and while some services only require a login to access the data, especially smaller companies or news organizations often require authentication (e.g., in the form of an ID copy) to proceed with data requests. Researchers are thus advised to carefully consider the instructions provided to participants and establish a support process for questions and consider potential biases in the recruitment (see Boeschoten et al., 2020).

Finally, while the data contained in DDPs can be extremely useful for academic research, they often do not include every aspect of someone's platform usage (e.g., it may include the comments made by a user, but not the content of the posts *seen* by a user), despite most legal provisions – and the GDPR in particular – requiring much wider access (Ausloos & Dewitte, 2018). Moreover, the provided data are often not structured in a way that can be readily used in analyses. Researchers are therefore advised to consider the meaningfulness of the data for their research design, creating an analysis plan before the data collection starts.

Despite the above challenges, we still strongly believe they are outweighed by the positive aspects of this method of digital data collection. The model adopted by OSD2F enables access to non-reactive digital data on a person-level that can be combined with self-reports. With OSD2F, we hope to facilitate detailed insights into interpersonal processes as well as the investigation of the digital infrastructures that underlie contemporary society more broadly, overcoming some of the many challenges to data access by academic research. Finally, by putting transparency, privacy, and individual agency at its core, OSD2F actively engages citizens in academic research and hopefully raises general awareness over the data collection practices of online platforms.

Author's note

Author names in alphabetical order.

Acknowledgements

This work was supported by the RPA Communication and its Digital Communication Methods Lab (digicomlab.eu) at the University of Amsterdam. DT received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 947695) for his contribution. JM received funding from the Dutch Science Foundation (Grant number: VI.Veni.191S.089). JA received funding from the Dutch Science Foundation (Grant number: VI.Veni.201R.096).

Notes

1. As was the case of the AdObserver (https://adobserver.org/), designed to only collect advertising-related information as participants were exposed to ads on Facebook. Tools like AdObserver, hence, can provide another excellent opportunity for researchers who want to give participants the possibility to donate their data. However, such real-time data donation tools are conceptually different from the donation of existing data, which we will discuss in the next section.
2. As providing an overview of the information contained in DDPs is outside of the scope of this paper (and may be outdated quickly), we advise interested researchers to download their own data from the relevant platforms to understand the breadth and the depth of the information available.
3. Other approaches to data donation currently exist or are in development. One example is the proof-of-concept PORT (Boeschoten et al., 2021), that focuses on the local extraction and analysis of DDPs in a centralized platform. Its approach may be seen as complementary to the options discussed here, yet it differs by its emphasis of *local* measurement of predefined concepts (instead of the donation of non-aggregated data).
4. Server library used is Quart.
5. https://github.com/uvacw/osd2f/
6. https://tortoise-orm.readthedocs.io/en/latest/
7. OSD2F is written with sample code for authentication using one cloud provider. This may differ depending on institutional requirements, and the researcher could set up secure direct access to the database if needed.
8. E.g., https://linktotheos2dfserver.eu/?sid=UNIQUEIDENTIFIER.
9. A more detailed analysis of the different legal, ethical and methodological considerations arising in this context, can be found in earlier work by one of the authors: Ausloos, Jef, and Michael Veale. 'Researching with Data Rights'. *Technology and Regulation*, 2020, 149–57.

10. Examples include the California Consumer Privacy Act ("right to know") or the Brazilian Lei Geral de Proteção de Dados (chapter III).
11. And, where the GDPR applies all the other information listed in Article 13 GDPR.
12. As with other open-source software, OSD2F is provided without warranty or liability by the developers.

References

Araujo, T., Wonneberger, A., Neijens, P., & Vreese, C. de. (2017). How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use. *Communication Methods and Measures*, *11*(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Ausloos, J., & Dewitte, P. (2018). Shattering one-way mirrors – data subject access rights in practice. *International Data Privacy Law*, *8*(1), 4–28. https://doi.org/10.1093/idpl/ipy001

Ausloos, J., & Veale, M. (2020). Researching with Data Rights. *Technology and Regulation*, *2020*, 136–157.

Baumgartner, S. E., Sumter, S. R., Petkevic, V., & Wiradhany. (2021). *Presenting a Novel Data Collection and Automated Processing Approach for iOS Smartphone Data*. 71st Annual ICA Conference, Virtual Conference.

Bodó, B., Helberger, N., Irion, K., Zuiderveen Borgesius, F., Moller, J., van de Velde, B., Bol, N., van Es, B., de Vreese, C., IViR (FdR), & Political Communication & Journalism (ASCoR, FMG). (2017). Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents. *Yale Journal of Law and Technology*, *19*, 133–180.

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). Digital trace data collection through data donation. *ArXiv:2011.09851* [*Cs, Stat*].

Boeschoten, L., Mendrik, A., van der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., & Oberski, D. (2022). Privacy preserving local analysis of digital trace data: A proof-of-concept. *Patterns,* 3(3). https://doi.org/10.1016/j.patter.2022.100444

Brinberg, M., Vanderbilt, R. R., Solomon, D. H., Brinberg, D., & Ram, N. (2021). Using technology to unobtrusively observe relationship development. *Journal of Social and Personal Relationships*, 02654075211028654. https://doi.org/10.1177/02654075211028654

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

Christner, C., Urman, A., Adam, S., & Maier, M. (2021). Automated Tracking Approaches for Studying Online Media Use: A Critical Review and

Recommendations. *Communication Methods and Measures*, 1–17. https://doi.org/10.1080/19312458.2021.1907841

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *0*(0), 1–15. https://doi.org/10.1080/1369118X.2019.1627386

Kobayashi, T., & Boase, J. (2012). No Such Effect? The Implications of Measurement Error in Self-Report Measures of Mobile Communication Use. *Communication Methods and Measures*, *6*(2), 126–143. https://doi.org/10.1080/19312458.2012.679243

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196. https://doi.org/10.1038/s41586-021-03660-7

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science (New York, N.Y.)*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Menchen-Trevino, E. (2016). *Web historian: Enabling multi-method and independent research with real-world web browsing history data*. https://doi.org/10.9776/16611

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2020). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication*, 2050157920959106. https://doi.org/10.1177/2050157920959106

Puschmann, C. (2019). Beyond the Bubble: Assessing the Diversity of Political Search Results. *Digital Journalism*, *7*(6), 824–843. https://doi.org/10.1080/21670811.2018.1539626

Reeves, B., Robinson, T., & Ram, N. (2020). Time for the Human Screenome Project. *Nature*, *577*(7790), 314–317. https://doi.org/10.1038/d41586-020-00032-5

Scharkow, M. (2016). The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data. *Communication Methods and Measures*, *10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Verbeij, T., Pouwels, J. L., Beyens, I., & Valkenburg, P. M. (2021). The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports*, *3*, 100090. https://doi.org/10.1016/j.chbr.2021.100090