Integrating Data Donations into Online Surveys

Mario Haim / Dominik Leiner / Valerie Hase*

Data donations represent a user-centered approach to data collection where researchers ask EU participants to exercise their right of access (GDPR) vis-à-vis intermediaries and to donate the digital trace data they receive to academic research. These data donations are often combined with survey data to gain deeper insights into the questions under investigation. Although initially promising, this process is complex for respondents and involves serious methodological, ethical, and legal challenges for researchers. A series of recently developed software solutions facilitate and streamline data donation studies. However, these stand-alone systems work separately from survey software. As a result, respondents typically face two platforms, one for the survey and one for the data donation. To facilitate their combination, we integrated two existing software solutions for online surveys (SoSci Survey) and data donations (OSD2F). We present our integrated solution and report on experiences with the approach from two exemplary studies.

Keywords: digital trace data, digital platforms, survey research, computational methods, GDPR, data collection

1. Introduction

The European Union's *General Data Protection Regulation* (GDPR) has created a legal basis for researchers to collect digital trace data from individuals. GDPR articles 15 (right of access by the data subject) and 20 (right to data portability), in particular, grant users the right to access personal data that intermediaries ("platforms") store about them. Intermediaries must provide users with structured and machine-readable access to their data. The core idea of data donation studies emerging from this legal shift, then, is to (1) inform individuals about this legal right, (2) ask them to request their data from intermediaries, (3) have them download their "Data Download Packages" (DDPs; Araujo et al., 2022, p. 375) and (4) donate their data to researchers.

Data donation studies build on a clear legal basis (Ohme & Araujo, 2022) and enable increased independence from intermediaries (Breuer et al., 2022). In a period of heightened barriers toward social media and the ways platforms store and use private data, data donations as a user-centered approach to the collection of digital trace data (Breuer et al., 2022) are the method of the hour that allows participants to retain sovereignty over their data against both intermediaries and researchers. Data donation studies, self-evidently, require informed consent and active cooperation from participants. Additionally, researchers should enable participants to inspect, filter, and delete data at various stages during the donation process. While it is crucial for research to retain access to digital trace data, another

* Prof. Dr. Mario Haim, Ludwig-Maximilians-Universität München, Institut für Kommunikationswissenschaft und Medienforschung, Akademiestr. 7, 80799 München, Deutschland, mario.haim@ifkw.lmu.de, https://orcid.org/0000-0002-0643-2299;
Dr. Dominik Leiner, Ludwig-Maximilians-Universität München, Institut für Kommunikationswissenschaft und Medienforschung, Oettingenstr. 67, 80538 München, Deutschland, dominik.leiner@ifkw.lmu.de, https://orcid.org/0000-0002-3862-3399;
Dr. Valerie Hase, Ludwig-Maximilians-Universität München, Institut für Kommunikationswissenschaft und Medienforschung, Akademiestr. 7, 80799 München, Deutschland, valerie.hase@ifkw.lmu.de, https://orcid.org/0000-0001-6656-4894.

 DOI: 10.5771/1615-634X-2023-1–2-130

outstanding advantage of data donations is the option to enrich the digital traces with survey data (Stier et al., 2020) on an individual basis.

However, recent data donation studies suggest considerable challenges (van Driel et al., 2022). These include, in particular, the sensitivity of obtained data (Boeschoten, Ausloos, et al., 2022), low response rates, and potential biases in acquired samples (Ohme et al., 2021; Pfiffner & Friemel, 2023). Regular changes to the structure of DDPs and a lack of coherence on how intermediaries provide data further complicate data collection (Breuer et al., 2022). Researchers thus face ethical and legal constraints, technical obstacles, high incentivization costs, and the challenge of non-representative findings.

2. Existing Research Software

Recently developed research software addresses these challenges by streamlining DDP handling, increasing response rates through user-friendly environments, and ensuring the anonymity of sensitive data through filtering and semi-automated anonymization. These projects include the *OSD2F* (Araujo et al., 2022) and *PORT* (Boeschoten, Mendrik, et al., 2022) systems originating in Amsterdam, as well as the *DDM* (Pfiffner et al., 2022) system developed in Zurich. While *PORT* and *DDM* are currently under development, *OSD2F* is already available as open source (licensed under GNU AGPL 3.0).

All systems represent stand-alone research software in that they are installed on a web server (on-premises) to enable data donations. *PORT* and *DDM* are conceptualized as platform solutions that allow multiple research projects within one installation, while each *OSD2F* installation is bound to an individual project. When running on-premises, all three systems offer researchers full control over sensitive data but require a great deal of technical knowledge for setup, maintenance, and data management.

The software packages offer limited functionality to pose questions to participants during the donation process but lack advanced survey features. Additional research software is necessary if researchers decide to integrate data donations into a larger research design, such as a survey, experimental variation, respondent recruitment through panels, or scheduled reminder emails about study participation. However, multiple online applications with various hyperlinks/URLs and different user interfaces may easily result in losing respondents along the way.

3. Integration of OSD2F with SoSci Survey

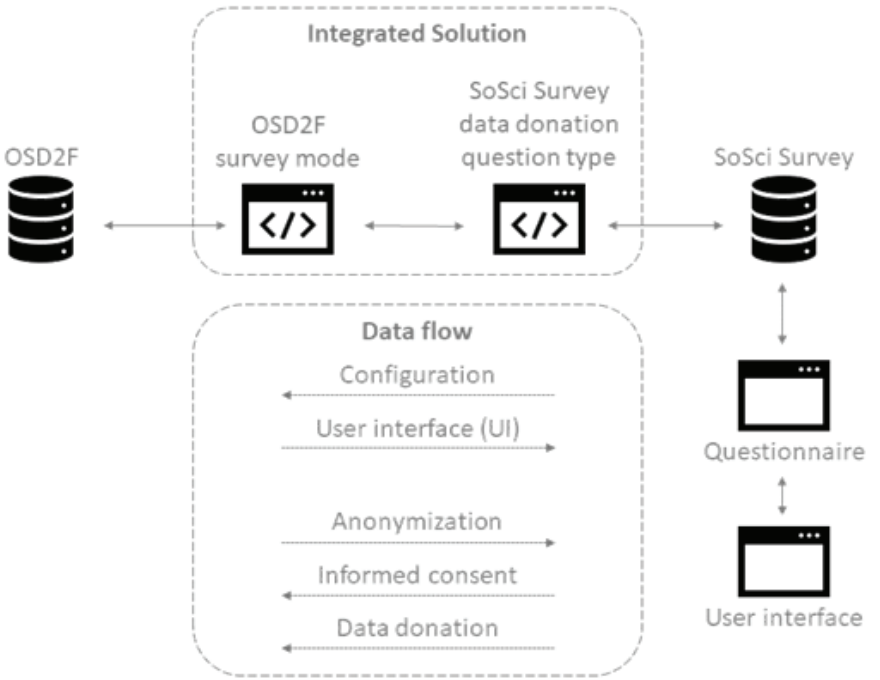
We propose integrating existing research software for data donations and online surveys to increase usability and decrease drop-outs. The proposed integration builds on *OSD2F* (Araujo et al., 2022) for data donations and the established web survey software *SoSci Survey* (Leiner, 2023). We implemented new features into each software to allow seamless integration (Fig. 1).

For *OSD2F*, we designed and developed a "survey mode," transforming the software from a stand-alone (browser front-end) to a supporting (API) system. As the developers of *OSD2F* system have moved on to build the *PORT* platform, our survey mode will live on as a fork to the seminal GitHub repository which itself will continue to receive maintenance updates. The survey mode is open to other software, its code has been shared open source under the same license as *OSD2F* (GNU AGPL 3.0), and extensive documentation is available in both our[1] and the seminal[2] GitHub repository.

1 https://github.com/datenfruehstueck/osd2f/blob/main/docs/survey.md.
2 https://github.com/uvacw/osd2f/tree/main/docs.

For *SoSci Survey*, we developed a new question type that uses the *OSD2F* survey mode, looping through data-donation-relevant parts of the *OSD2F* front-end into the questionnaire. The integration routes only the front-end through *SoSci Survey* while the data flows directly back to *OSD2F*. In its most recent version, *SoSci Survey* includes the new question type by default so that no additional installation is necessary. Since data donations are managed in *OSD2F* while survey data remains in *SoSci Survey*, the genuine strengths of both systems are retained and combined.

Figure 1: Schematic representation of the integration process



Both *OSD2F* and *SoSci Survey* can be installed on self-maintained servers (on-premises), although there are also cloud options (software-as-a-service) for *SoSci Survey*. Requirements are comparably low for both *OSD2F* (Python, SQLite database) and *SoSci Survey* (PHP, MySQL database). However, we recommend running both via hosting providers to simplify the setup process. For *OSD2F*, the GitHub repository includes scripts for container deployment, for example, via an Azure or a Docker instance. Our GitHub repository includes additional details on *OSD2F*'s survey mode and its technical specification as well as its installation and deployment, either via *Docker* or through a webserver such as *nginx*. As this instance will also store data donations, the location of the hosting providence may be of legal relevance—EU-based studies may prefer a hosting provider operating in the EU. Similarly, studies in the EU can choose a European software-as-a-service solution for *SoSci Survey* (e.g., www.soscisurvey.de).

4. Setting Up the Integrated Research Solution

Our proposed solution runs along five steps, divided into a setup phase (steps 1 and 2) and a runtime phase (steps 3 to 5).

4.1 Setup Phase

The setup phase begins after both systems are running and a researcher adds a question based on the new data donation question type to a questionnaire in *SoSci Survey*. In step 1, the researcher defines settings for the data donation question and its appearance, including the *OSD2F* server's URL, the configuration on which DDPs to expect, and what anonymizer to apply to uploaded data (e.g., replacing usernames with a generic "<user>"). A JSON object similar to *OSD2F*'s default "upload configuration" represents the DDP configuration. It is maintained through a respective input field in the *SoSci Survey* data donation question type (Fig. 2). *SoSci Survey* then reaches out to *OSD2F* to install (or update) the *OSD2F* configuration. We strongly recommend using an HTTPS (SSL) URL for the *OSD2F* instance to encrypt all communication. In step 2, *OSD2F* responds to this installation request, providing *SoSci Survey* with the necessary HTML, CSS, and JavaScript code for the user interface ("UI") to display the data donation input screen. *SoSci Survey* stores this code to embed the donation form in the questionnaire.

4.2 Runtime Phase

The runtime phase is initiated when a respondent fills out the questionnaire and arrives at the page displaying the data donation input screen (Fig. 3). This interface includes visible elements and the *OSD2F* scripts that run in the respondent's browser. In the background, *SoSci Survey* provides the interface with the interview case number to merge data donations and survey data later.

Step 3 is triggered when the respondent uses the interface to select one or more DDPs. The interface will preprocess the DDPs, filter out unwanted files, and automatically extract individual data items. In the background, it requests anonymizers from the *OSD2F* server while allowing respondents to inspect and optionally delete items (a screenshot from our German exemplary study is provided in Fig. 4) and eventually requests their informed consent for the donation. When respondents consent, the interface will transfer (step 4) the individual data items directly to the *OSD2F* instance, including the interview case number. The HTTPS endpoint, again, ensures an encrypted connection.

As per its stand-alone behavior, *OSD2F* acknowledges the upload with status information (step 5). The interface will display status feedback if necessary and redirects the information to *SoSci Survey* to record the number of donated and deleted items in the survey dataset. In addition, the interface will signal *SoSci Survey* to continue with the remaining questionnaire, and the runtime phase ends.

4.3 Data Management

Data management requires pulling datasets out of two separate systems. Here, we mainly rely on R: Researchers can download survey data from *SoSci Survey* as a file, via direct access (API), or via the "soscisurvey" R package (Unkel, 2019). For *OSD2F*, the SQLite database file needs to be downloaded, for example, via SFTP. Researchers can then unpack the data donation database using the *DBI* R package, *RSQLite*, and *tidyverse*. A dbConnect() function call to the database file is necessary before using dbGetQuery() to collect all data

Figure 2: Screenshot of the SoSci Survey Data Donation Question Type

donations ("SELECT * FROM submissions"). A final as_tibble() function call transforms the data donations into a familiar tabular format. Lastly, researchers can merge both datasets through the respondents' anonymous case numbers, that is *SoSci Survey*'s "CASE" and *OSD2F*'s "submission_id" variables, for example, through the join() functions from the *tidyverse* package.

Figure 3: Preview of the Data Donation Input Screen Shown in SoSci Survey

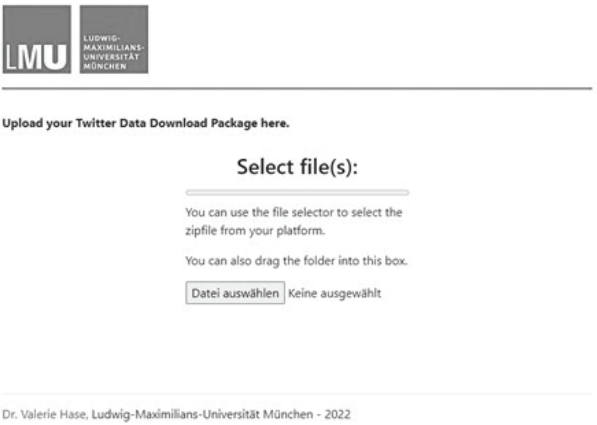
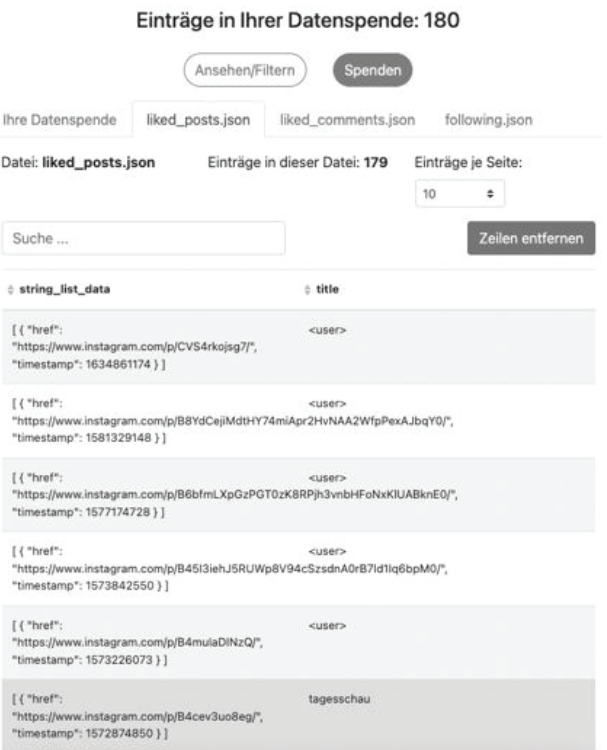


Figure 4: Screenshot of the Inspection Screen for Users to Filter Anonymized Data from a German Exemplary Study



5. Exemplary Studies

We tested our integrated approach in two studies. Both invited participants to take part in a survey on their digital news use and, subsequently, to donate DDPs from up to four intermediaries (Facebook, Instagram, Twitter, and YouTube).

The studies differed in their recruitment. Students personally approached participants for study 1 (N = 345) and offered them face-to-face support for retrieving and uploading DDPs. Participants for study 2 (N = 2.039) were recruited through an access panel (SoSci Panel, Leiner, 2016); the instructions explained DDP handling in detail but did not explicitly offer any support options. As such, study 2 particularly highlights the strength of our integrated approach. Our approach grants researchers access to all features known from survey platforms, such as multilingual surveys, connecting to an access panel, sending reminders for participation, or distributing incentives. Participants partake in the survey and the data donation through a single platform. Moreover, survey platforms such as *SoSci Survey* allow for pausing while filling out the questionnaire—for example, to request DDPs and wait for the platform (often for days) to provide the DDPs. The integration functioned without any noticeable setbacks in both studies. Actual shares of donors varied between study 1 (n = 69; 20%) and study 2 (n = 245; 12%). An open-ended question to those deciding against data donation mainly revealed privacy concerns and little use of requested platforms as reasons. Participants rarely mentioned technical failures or incoherent user experience.

6. Conclusion

Facing ever-declining response rates (de Leeuw et al., 2018), survey research must minimize the obstacles to respondents' participation in scientific research. Data donation studies offer solutions to several challenges social-media research has faced in recent years, especially the dependency on intermediaries and measurement error in self-reported data on media use. However, data donations also introduce additional obstacles, such as increased burdens for gathering representative samples due to low and often biased response rates. Our proposed integration of *OSD2F* with *SoSci Survey* aims to facilitate data donation studies, both for participants and researchers. For participants, the integration eliminates the need to switch platforms when participating in data donation studies. Researchers can rely on the strengths of both of these research software solutions to collect digital trace data and survey responses.

Literature

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

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A Framework for Privacy Preserving Digital Trace Data Collection through Data Donation. *Computational Communication Research*, *4*(2), 388–423. https://doi.org/10.5117/CCR2022.2.002.BOES.

Boeschoten, L., Mendrik, A., van der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., & Oberski, D. L. (2022). Privacy-Preserving Local Analysis of Digital Trace Data. A Proof-of-Concept. *Patterns*, *3*(3), 100444. https://doi.org/10.1016/j.patter.2022.100444.

Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2022). User-Centric Approaches for Collecting Facebook Data in the "Post-API Age". Experiences from Two Studies and Recommendations for Future Research. *Information, Communication & Society*, *0*(0), 1–20. https://doi.org/10.1080/1369118X.2022.2097015.

de Leeuw, E., Hox, J., & Luiten, A. (2018). International Nonresponse Trends Across Countries and Years. An Analysis of 36 Years of Labour Force Survey Data. *Survey Methods: Insights from the Field (SMIF)*. https://doi.org/10.13094/SMIF-2018-00008.

Leiner, D. J. (2016). Our Research's Breadth Lives on Convenience Samples. A Case Study of the Online Respondent Pool "SoSci Panel." *Studies in Communication and Media*, *5*(4), 367–396. https://doi.org/10.5771/2192-4007-2016-4-367.

Leiner, D. J. (2023). *SoSci Survey* (3.4.11) [Computer software]. https://www.soscisurvey.de [16.03.2023].

Ohme, J., & Araujo, T. (2022). Digital Data Donations. A Quest for Best Practices. *Patterns*, *3*(4), 100467. https://doi.org/10.1016/j.patter.2022.100467.

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile Data Donations. Assessing Self-Report Accuracy and Sample Biases with the iOS Screen Time Function. *Mobile Media & Communication*, *9*(2), 293–313. https://doi.org/10.1177/2050157920959106.

Pfiffner, N., & Friemel, Thomas. N. (2023). Leveraging Data Donations for Communication Research: Exploring Drivers Behind the Willingness to Donate. *Communication Methods and Measures*, 1–23. https://doi.org/10.1080/19312458.2023.2176474.

Pfiffner, N., Witlox, P., & Friemel, T. N. (2022). *Data Donation Module (Version 0.1.26)* [Computer software]. https://datadonation.uzh.ch/ [16.03.2023].

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating Survey Data and Digital Trace Data. Key Issues in Developing an Emerging Field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669.

Unkel, J. (2019). *SoSciSurvey*. https://github.com/joon-e/soscisurvey [16.03.2023].

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures*, *16*(4), 266–282. https://doi.org/10.1080/19312458.2022.2109608.