




Port: A software tool for digital data donation

Laura Boeschoten ¹[¶], Niek C. de Schipper ², Adriënnë M. Mendrik ³,
Emiel van der Veen ³, Bella Struminskaya ¹, Heleen Janssen ², and
Theo Araujo ²

1 Utrecht University, The Netherlands 2 University of Amsterdam, The Netherlands 3 Eyra Leap B.V., the Netherlands [¶] Corresponding author

DOI: [10.21105/joss.05596](https://doi.org/10.21105/joss.05596)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: Øystein Sørensen  

Reviewers:

- [@kaustavbhattacharjee](#)
- [@leonardojaneis](#)

Submitted: 03 June 2023

Published: 03 October 2023

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

Recently, a new workflow has been introduced that allows academic researchers to partner with individuals interested in donating their digital trace data for academic research purposes (Boeschoten, Ausloos, et al., 2022). In this workflow, the digital traces of participants are processed locally on their own devices in such a way that only the subset of participants' digital trace data that is of legitimate interest to a research project are shared with the researcher, which can only occur after the participant has provided their informed consent.

This *data donation workflow* consists of the following steps: First, the participant requests a digital copy of their personal data at the platform of interest, such as Google, Meta, Twitter and other digital platforms, i.e., their *Data Download Package* (DDP). Platforms, as data controllers, are required as per the European Union's General Data Protection Regulation (GDPR) to share a digital copy with each participant requesting such a copy. Second, they download the DDP onto their personal device. Third, by means of *local processing*, only the data points of interest to the researcher are extracted from that DDP. Fourth, the participant inspects the extracted data points after which the participant can consent to donate. Only after providing this consent, the donated data is sent to a storage location and can be accessed by the researcher, which would mean that the storage location can be accessed for further analysis.

In this paper, we introduce Port. Port is a software tool that allows researchers to configure the local processing step of the data donation workflow, allowing the researcher to collect exactly the digital traces needed to answer their research question. When using Port, a researcher can decide:

- Which digital platforms are investigated;
- Which digital traces are collected;
- How the extracted digital traces are visually presented to the participant;
- What is communicated to the participant.

Statement of need

In our everyday lives, we leave more and more digital traces behind on digital platforms: for example, by liking a post on Instagram or sending a message via WhatsApp; when we tap our electronic card on public transportation or complete an online banking transaction. The promise of digital humanities and computational social science is that researchers can utilize these digital traces to study human behavior and social interaction at an unprecedented level of detail (King, 2011).

However, while the amount of digital trace data increases, most are closed off in proprietary archives of commercial corporations, with only a subset being available to a small set of

researchers at a platform's discretion, through initiatives such as Social Science One (King & Persily, 2020), or through increasingly restricted and opaque APIs (Brunns, 2019; Freelon, 2018; Perriam et al., 2020).

An alternative approach to gain access to digital traces is enabled thanks to the GDPR's right to data access and data portability (Ausloos & Veale, 2021). Thanks to this legislation, all data processing entities are required to provide citizens a digital copy of their personal data upon request in, where that is appropriate, electronic form. We refer to these pieces of personal data as *Data Download Packages* (DDPs).

This legislation allows researchers to invite participants to share their DDPs. A major challenge is, however, that DDPs potentially contain very sensitive data. Conversely, often not all data is needed to answer the specific research question. To tackle these challenges, Boeschoten, Ausloos, et al. (2022) developed an alternative workflow: First, the participant requests their personal DDP at the platform of interest. Second, they download it onto their own personal device. Third, by means of *local processing*, only the features of interest to the researcher are extracted from that DDP. Fourth, the participant inspects the extracted features after which they can choose what they want to donate (or decline to donate). Only after selecting the data for donation and clicking the button *donate*, the donated data is sent to a storage location and can be accessed by the researcher. See Figure 1 for an overview of these steps.

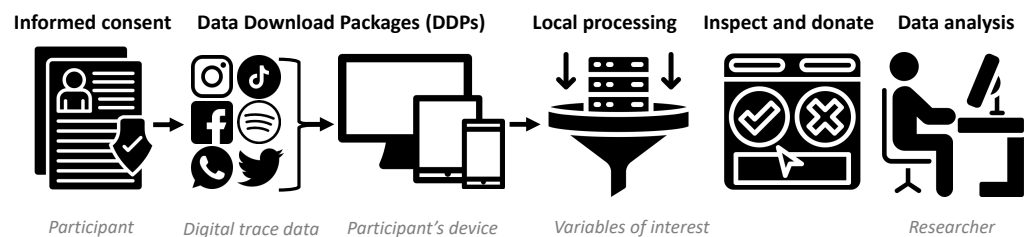


Figure 1: An overview of the participant's data donation flow as presented by Boeschoten, Ausloos, et al. (2022).

In recent years, researchers have implemented the local processing step in multiple ways. Studies by van Driel et al. (2022) and Kmetty & Németh (2022) used an approach where DDPs were donated directly, so without a local processing step taking place. However, extensive de-identification procedures were in place to guarantee participant's privacy (see e.g. Boeschoten et al. (2021)). In addition, multiple apps have been developed to enable this local processing step. For example, the app OSD2F (Araujo et al., 2022) allows to select .json files within the DDP and performs certain de-identification steps prior to donation. Alternatively, the app Port (Boeschoten, Mendrik, et al., 2022) allows for more rigorous data processing, as researchers can write a custom Python script that locally performs any requested processing task. While this approach undoubtedly offers greater flexibility, it concurrently places a larger level of responsibility on the researcher. Other apps have been documented in academic literature. However, the feasibility of their reuse for researchers outside their respective institutions or for other platforms as intended is more ambiguous. Examples of such apps are Designerly Data Donation (Ortega et al., 2021), the Data Donation Module (Piffner et al., 2022), the Social Media Donator (Zannettou et al., 2023) and WhatsR (Kohne, 2023).

In this paper, we introduce a new version of Port. It is open-source and allows for researchers to fully configure their own data donation study. It creates an app that guides participants through the data donation steps. Researchers can tailor this app to the DDP of their platform of interest and process these in their desired ways. In addition to local processing, key features from the Open-Source Data Donation Framework (OSD2F) are also integrated, allowing participants to decide per data instance whether they want to exclude it from being donated.

Note that researchers always ask permission from their own Ethical Review Boards (ERBs) and Data Protections Officers (DPOs), and that using Port does not dismiss researchers from these obligations. The purpose of Port is to enable researchers to access platform user data with a GDPR compliant approach.

Which digital platforms are investigated?

Port is a tool that allows researchers to collect digital traces through donation of DDPs. In practice, this means that Port can be configured to process DDPs from any data controller, i.e., any legal entity that processes personal data. However, collection of digital traces through data donation using Port can only be a viable approach for data collection if the platform acting as data controller meets certain criteria.

First, in order for data donation research to occur, a platform must comply with the individual data access request that was submitted to it, meaning that platform compliance with the GDPR is a condition sine qua non for effective data donation research. Second, the process to request a copy of one's personal data should be standardized to a certain extent such that researchers can provide study participants with instructions on how to do this. Third, the file format of the DDP should ideally have a certain level of standardization as well. It is not possible to plan the procedure or extraction of data from the DDP if it is unknown to the researcher where the data of interest can be found within the DDP.

How are digital traces extracted?

Port consists of two distinct elements, which are both fully controlled by a Python script that runs locally in the browser of the participant. This Python script is specifically tailored for each data donation study. The first element is the data donation study flow. The goal of this part of the Python script is to provide explanations or instructions to the participant at various steps of the flow. The second element is the data extraction process. The goal of this part is to make sure that only the digital traces that are of interest to the researcher are extracted from the DDPs and that were agreed upon by the participant.

To run a custom Python script, Port makes use of Pyodide ([The Pyodide development team, 2021](#)). Pyodide is a Python distribution for the browser based on WebAssembly ([WebAssembly, 2021](#)).

Running the custom Python script using Pyodide in the browser of a participant works as follows:

- The Python script starts and begins to run synchronously, until:
 1. The script reaches a Python class resembling a UI element that should be shown on screen, a React component ([React, 2022](#)).
 2. The script yields and communicates with the app which UI should be rendered on screen.
 3. The participant interacts with the UI element.
 4. The outcome of the interaction is passed back to the Python script and can be handled accordingly.
- Steps 1 through 4 are repeated until the end of the Python script.

A Python script for a data donation study typically contains the following steps:

- The welcome screen for the data donation process is shown.
- The participant is asked to submit their DDP.
- The input is validated.
- The digital traces of interest to the researcher are extracted from the DDP.
- The extracted digital traces are placed in a table.

- The table is rendered on screen.
- The participant clicks on the 'Yes, donate' or 'No' button.
- The closing screen is shown.

A. Snippet of the Google Semantic Location History (GSLH) DDP

```

2016_NOVEMBER - Notepad
File Edit Format View Help
{
  "endLocation" : {
    "latitudeE7" : 520893191,
    "longitudeE7" : 51101691,
    "placeId" : "ChIJeb1WZV1vxkcRbk1MYz1wjbq",
    "address" : "Stationshal 12, 12\n3511 CE Utrecht\nNetherlands",
    "name" : "Utrecht Centraal",
    "locationConfidence" : 100.0
  },
  "duration" : {
    "startTimestampMs" : "1478114254623",
    "endTimestampMs" : "1478114520071"
  },
  "confidence" : "LOW",
  "activities" : [ {
    "activityType" : "IN_TRAIN",
    "probability" : 0.0
  }, {
    "activityType" : "CYCLING",
    "probability" : 38.266496335674674
  }, {
    "activityType" : "IN_PASSENGER_VEHICLE",
    "probability" : 1.8217724982410624
  }
]
}
Ln 3045, Col 9 | 140% | Unix (LF) | UTF-8

```

B. Locally processed to extract the distance travelled and duration per activity type

Cycling

		Duration (hours)	Distance (km)
Year	Month		
2016	11	5.32	91.49
	12	12.98	199.51
2017	1	8.60	121.26
	2	13.40	308.93
	3	12.43	198.12

Figure 2: A. shows an example of a location visit in the Google Semantic Location History (GSLH) Data Download Package (DDP). B. shows how this DDP was processed into a frequency table presenting the distance and duration per activity type per month.

The benefit of having a Python script running inside the browser is that the researcher has familiar tools to design the extraction process in such a way that the privacy of the participants is preserved as much as possible. For this purpose, the researcher can make use of two important features. First, besides extracting digital traces from the DDP, it is also possible to further process these to better match the research question. [Figure 2](#) shows an example of raw Google Semantic Location History (GSLH) data that is locally processed to only extract the duration and distance of the various activities tracked by GSLH per month.

A. Snippet of a WhatsApp chat DDP

```

chat - Notepad
File Edit Format View Help
[21/01/2022, 17:55:56] You created group "WhatsApp project"
[21/01/2022, 17:55:56] WhatsApp project: Messages and calls are end-to-end
[21/01/2022, 18:02:32] Research Colleague 1: Hi!
[21/01/2022, 18:02:53] Laura: ☺
[21/01/2022, 18:03:18] Laura: Hi all
[21/01/2022, 18:03:34] Research Colleague 1: How are you?
[21/01/2022, 18:03:46] Laura: I'm good! How are you?
[21/01/2022, 18:15:12] Research Colleague 1: 🙄
[21/01/2022, 19:56:11] Research Colleague 2 joined using your invite
[21/01/2022, 21:47:27] Research Colleague 1: Welcome!
[22/01/2022, 13:18:31] Research Colleague 2: Thanks! Is this the group wh
[22/01/2022, 13:24:17] Research Colleague 1: 🙄
[22/01/2022, 13:24:29] Research Colleague 1: You can of course!
[22/01/2022, 13:51:43] Laura: You are also allowed to talk about other th
[22/01/2022, 13:52:09] Laura: image omitted
[22/01/2022, 19:38:02] Research Colleague 2: That looks really good!
[17/02/2022, 13:25:15] Research Colleague 2 added Research Engineer 2
[16/03/2022, 15:16:30] Research Engineer 2 : Hi Laura, I am not the admin!
[16/03/2022, 15:17:11] Research Engineer 2 : I will share with you her pi
[16/03/2022, 15:17:32] Laura: I made you admin
[16/03/2022, 15:17:45] Research Engineer 2 : Great, thanks 🙄
[16/03/2022, 15:18:27] Research Engineer 2 added Research Engineer 1
[16/03/2022, 15:20:25] Laura: Hi shival!
[16/03/2022, 15:25:39] Research Engineer 1 : Hi ☺
Ln 16, Col 69 100% Windows (CRLF) UTF-8

```

B. Names are locally extracted and prompted on screen. The participant selects their own name.

C. The participant is identified in the extracted data

Select username

Please indicate which username is yours. Note that names and phone numbers are not stored, but only used to extract relevant information from the chat file.

- Laura
- Research Engineer 1
- Research Engineer 2
- Research Colleague 1
- Research Colleague 2
- Mijn naam of telefoonnummer staat er niet tussen

Dit bent u

	Omschrijving	Gegevens
0	Aantal woorden	131
1	Aantal berichten	31
2	Datum eerste bericht	2022-01-21 18:02:00
3	Datum laatste bericht	2022-03-24 21:03:00
4	Aantal websites	3
5	Aantal foto's en bestanden	1
6	Aantal localities	2
7	Wie reageert het meest op deze deelnemer?	Deelnemer 1
8	Op wie reageert deze deelnemer het meest?	Deelnemer 1

Figure 3: A. shows an example of a WhatsApp chat Data Download Package (DDP). B. shows how first only the usernames of the members in this chat were locally extracted. The participant can select their own username from this list. C. shows how this DDP was processed into a frequency table presenting among other things how often the members respond to each other's actions. Here, the participant is identified, the others receive anonymous labels. Note that the output is presented in Dutch.

Second, a local interaction between the participant and the DDP can be added as well, so that the participant can provide context to the data. Figure 3 shows an example using a DDP of a WhatsApp group chat. The Python script extracts the names of all people in the chat, which are then presented to the participant in a way that they can select their own name. This functionality can for example be used to identify the participant within their WhatsApp group in order to extract the messages that were written by the participant and discard all others in the group chat, or to count the number of messages to and from the participant. This functionality allows for the preservation of the privacy of other people in the group chat by asking feedback from the participant, since these people have not consented to the use of their data.

How are the extracted digital traces visualized?

A. Data extracted from a Twitter DDP

Twitter

Determine whether you would like to donate the data below. Carefully check the data and adjust when required. With your donation you contribute to the previously described research. Thank you in advance.

Zip file contents

< 1 2 3 4 5 6 7 > 9/9 pages Search

filename	compressed size	size
data/	2	0
data/README.txt	10003	40549
Your archive.html	730	1432
assets/	2	0
assets/images/	2	0
assets/images/groupAvatar.png	703	1354
assets/images/favicon.ico	486	481

Adjust No adjustments

Download

- Check the email that you received from Twitter
 - Click on the download link and store the file
 - Choose the stored file and continue
- [Click here](#) for more extensive instructions

Do you want to donate the above data?

B. Delete rows from the extracted data prior to donation

Twitter

Determine whether you would like to donate the data below. Carefully check the data and adjust when required. With your donation you contribute to the previously described research. Thank you in advance.

Zip file contents

< 1 2 3 4 5 6 7 > 9/9 pages Search

filename	compressed size	size
<input type="checkbox"/> data/	2	0
<input type="checkbox"/> data/README.txt	10003	40549
<input type="checkbox"/> Your archive.html	730	1432
<input type="checkbox"/> assets/	2	0
<input checked="" type="checkbox"/> assets/images/	2	0
<input type="checkbox"/> assets/images/groupAvatar.png	703	1354
<input type="checkbox"/> assets/images/favicon.ico	486	481

Adjust No adjustments

Download

- Check the email that you received from Twitter
 - Click on the download link and store the file
 - Choose the stored file and continue
- [Click here](#) for more extensive instructions

Do you want to donate the above data?

Figure 4: The top image shows an example of data that is extracted from a YouTube DDP, and is presented to a participant prior to providing consent. The participant can click on the 'adjust' button, after which rows can be selected for deletion (see bottom image).

After data extraction and potential further processing (as for example shown in [Figure 2](#)), the data is shown on screen for the participants to review, so that they can determine whether to donate the data. The data is shown to provide participants insight into what they share exactly, in order for them to provide a truly informed consent when deciding to donate this data to the

researcher. This visualization step also provides the participant with more autonomy over what is shared, as they can select specific data instances and delete them prior to donation (see [Figure 4](#)). Providing participants with this option is particularly interesting when working with sensitive data, such as text messages. Researchers that receive the donated data are informed by Port that data was deleted, but not which data was deleted. Custom user interface elements could be developed to allow for other types of interactions, such as labeling the data, or to present the data in other formats, such as in histograms, if suitable.

What is communicated to the participant?

Where a researcher invites participants for a data donation study, they may communicate their intentions to inform the individual participants. For example, researchers can generally inform participants in a more generic privacy policy about the purpose of the study or about the instructions on how to request and download the DDP of interest. Yet, to obtain unambiguous, specific and informed consent from individual participants, a researcher's consent form should indicate the specific purposes of the processing for which the use of a participant's personal data is intended. To communicate this information for a specific data donation study, all text that is prompted on screen can be adjusted. Currently, two languages (Dutch and English) are supported. There is room to link to external documents, which we have used in multiple studies to refer to the privacy policy and a document with data request and download instructions. Finally, Port collects paradata such as time stamps, information on clicks and navigation during the donation process. This paradata can be used to monitor if the information provided to the participants is clear or if there are problems with particular aspects.

Conclusion

To summarize, by utilizing GDPR's right of data access, data donation is a promising new approach to collect digital trace data for research purposes. The data donation workflow as introduced by Boeschoten, Ausloos, et al. (2022) introduces the idea of locally processing the obtained digital traces at the device of a participant as such that only the digital traces that are of legitimate interest to the researcher are shared. The software introduced in this paper, Port, allows a research to configure a custom data donation study. The research can decide: Which digital platform to investigate, which digital traces to collect, how to present the digital traces to the participant, and what to communicate to the participant throughout this process. These functionalities make Port a generic and useful tool for any researcher interested in collecting digital traces for research purposes.

Acknowledgements

The development of Port was partly made possible by the [Platform Digitale Infrastructuur SSH](#) in the Netherlands ("Digital Data Donation Infrastructure (D3I)") and in-kind contribution of Eyra Leap B.V.

References

- Araujo, T., Ausloos, J., Atteveldt, W. van, 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/10.31235/osf.io/xjk6t>
- Ausloos, J., & Veale, M. (2021). Researching with data rights. *Technology and Regulation*, 2020, 136–157. <https://doi.org/10.26116/techreg.2020.010>

- 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., Veen, E. van der, 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>
- Boeschoten, L., Voorvaart, R., Van Den Goorbergh, R., Kaandorp, C., & De Vos, M. (2021). Automatic de-identification of data download packages. *Data Science*, 4(2), 101–120. <https://doi.org/10.3233/ds-210035>
- 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>
- Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, 35(4), 665–668. <https://doi.org/10.31235/osf.io/56f4q>
- King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018), 719–721. <https://doi.org/10.1126/science.1197872>
- King, G., & Persily, N. (2020). A new model for industry–academic partnerships. *PS: Political Science & Politics*, 53(4), 703–709.
- 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. <https://doi.org/10.1177/07591063211061754>
- Kohne, J. (2023). “WhatsR - an r-package for parsing, anonymizing and visualizing exported WhatsApp chat logs. (Date accessed: 30.07.2023). <https://doi.org/10.5281/zenodo.7875622>
- Ortega, A. G., Bourgeois, J., & Kortuem, G. (2021). Towards designerly data donation. *UbiComp/ISWC 2021 - Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, 496–501. <https://doi.org/10.1145/3460418.3479362>
- Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. *International Journal of Social Research Methodology*, 23(3), 277–290. <https://doi.org/10.1080/13645579.2019.1682840>
- Pfiffner, N., Witlox, P., & Friemel, T. N. (2022). *Data donation module*. (Date accessed: 11.08.2023). <https://github.com/uzh/ddm>
- React. (2022). React. In *React*. (Date accessed: 11.08.2023). <https://react.dev/>
- The Pyodide development team. (2021). *Pyodide/pyodide* (Version 0.23.0). (Date accessed: 11.08.2023). <https://doi.org/10.5281/zenodo.5156931>
- 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/10.31219/osf.io/krqb9>
- WebAssembly. (2021). WebAssembly. In *WebAssembly*. (Date accessed: 11.08.2023). <https://webassembly.org/>
- Zannettou, S., Nemeth, O.-N., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., & Roesner, F. (2023). *Leveraging rights of data subjects for social media analysis: Studying TikTok via data donations*. <http://arxiv.org/abs/2301.04945>