A short course on how to prepare a data donation study using Port

Organizers.

Name: Laura Boeschoten
Affiliation: Utrecht University, Department of Methodology and Statistics
Email address: l.boeschoten@uu.nl
Bio: Laura Boeschoten works as an assistant professor at the Department of Methodology and Statistics at Utrecht University. She is the project lead of D3I, a consortium of researchers, research engineers and software developers in the Netherlands that work on the development of a Digital Data Donation Infrastructure (D3I). In addition, she focuses on integrating D3I within the Dutch national research infrastructure ODISSEI. Her research focuses on the various methodological challenges that researchers face when conducting data donation studies, both in terms of measurement quality and representation, and she often assists researchers in conducting data donation studies. Previously, she has worked as a postdoctoral researcher on the project 'Valid measures derived from incidental data' led by Daniel Oberski at Utrecht University, and has obtained her PhD in 2019 entitled 'Consistent estimates based on a mix of administrative data and surveys', which was a joint project between Tilburg University and Statistics Netherlands.

Name: Niek de Schipper
Affiliation: University of Amsterdam, Amsterdam School of Communication Research
Email address: n.c.deschipper@uva.nl
Bio: Niek de Schipper works as a Research Engineer at the University of Amsterdam, where he contributes to projects on digital data donation. De Schipper is an integral part of the Digital Data Donation community, helping many researchers in conducting their data donation studies. De Schipper has obtained his PhD in 2021 at Tilburg University in the Methods and Statistics department.

Topic.

Recently, a new workflow has been introduced that allows researchers to partner with individuals interested in donating their digital trace data for academic purposes. 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, X 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, study participants download the DDP onto their personal device. Third, by means of local processing, only the datapoints of interest to the researcher are extracted from that DDP. Fourth, the participant inspects the extracted datapoints 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 for further analysis.

Rationale.

Researchers from the computational social science and digital humanities community often struggle with the lack of access to data about online behavior. This challenge is even more pressing now that several APIs are closing. At the same time, in our everyday lives, we

as individuals 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. In summary, 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, or through increasingly restricted and opaque APIs.

This tutorial helps researchers understand and deploy an alternative to circumvent these challenges. This alternative approach to gain access to digital traces is enabled thanks to the GDPR's right to data access and data portability and similar legislation in other countries. As a result, all data processing entities are required to provide citizens a digital copy of their personal data upon request in 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, an alternative workflow has been developed: 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 and be used for further analysis.

After having participated in this tutorial, attendees will know what designing a data donation study entails and what important aspects should be considered. Attendees will learn about the different types of study designs in which data donation can be incorporated. Furthermore, attendees will learn how to configure their own data donation study using the open-source software Port and how to write their own Python scripts used for the extraction of digital trace data.

Format.

This course consists of four parts:

1. **Introduction to data donation.** We introduce the general concept of data donation. We illustrate the importance of data donation as an approach to collect digital trace data and introduce the basic principles of the data donation study. We provide examples of research questions for which data donation was used and discuss key methodological considerations when designing a data donation study.
2. **Prepare a data donation study design.** Attendees get the opportunity to configure their own data donation study. For this, attendees make use of the software that we have developed for conducting data donation studies: Port. We focus on study design elements such as:
   - How to recruit your research participants?
   - What other data sources do you want to combine with the digital trace data?
   - How to facilitate the data access request procedure of your research participants?

   We provide our considerations, materials and templates from an exemplary study design using WhatsApp data, and also guide the attendees in designing their own studies focusing on digital trace data from various digital platforms.
3. **Determine which digital traces to collect.** Port, is a generic tool. This means that it can be used to collect digital traces from any platform of interest (as long as the platform allows for its users to perform data access requests). However, in most cases, not all data that can

be collected of a specific platform actually need to be collected in order to answer a specific research question. In addition, collecting more data than needed to answer a research question can be considered ethically questionable.

In order to minimize the amount and type of digital traces that are collected during a data donation study, Port allows the researcher to develop a custom Python script that determines exactly which digital traces are extracted from a DDP and which traces are not collected. This Python script also provides the opportunity to facilitate a local interaction between the research participant and their DDP.

We provide the attendees with different alternative versions of Python scripts that illustrate how different data can be collected from the same platform (again with WhatsApp as a use case) and also facilitate that attendees can start writing their own custom Python scripts for their own research questions and platforms of interest.

4. **Deploying your data donation study.** In order to field your data donation study, three aspects are important to consider. First, the study needs to be hosted somewhere, such that it can be accessed by the research participants. Second, research participants need to be directed towards the study. This can, for example, be through an online access panel, which can provide the weblinks themselves, or the participants can be redirected through the panel software. Third, the researcher needs to configure where the data is stored once the participant clicks the 'donate' button. We provide the attendees with an overview of the possibilities and most important considerations, we explain what kind of solutions we have currently available and provide links to where the attendees can find detailed information on how to use the existing solutions.