Patterns

CellPress
OPEN ACCESS

Preview

Digital data donations: A quest for best practices

Jakob Ohme[1,2,*] and Theo Araujo[2]
[1]Weizenbaum Institute for the Networked Society, Freie Universität Berlin, Berlin, Germany
[2]Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, the Netherlands
*Correspondence: j.ohme@fu-berlin.de
https://doi.org/10.1016/j.patter.2022.100467

This preview article discusses PORT—a data donation software newly developed by Boeschoten et al.—toward the background of three core data donation principles: privacy protection, meaningful data extraction, and securing user agency.

The increasing emphasis on data rights—such as those mandated by the European Union's General Data Protection Regulation (GDPR) —means that users can now request and download the data that digital platforms and other companies collect about them, often in the format of data download packages (DDPs).[1] Access to these data—with individuals voluntarily donating their DDPs to academic research—may open a "treasure trove for social scientists"[2] and create a unique opportunity to explore crucial research questions that can best be answered with access to digital traces of individuals. This is particularly pressing if the social sciences are to derive meaningful measures of human behavior in increasingly algorithmically infused societies (for an overview, see Lazer et al., 2021[3] and Wagner et al., 2021[4]).

As research making use of data donation begins to gain traction in the social sciences with different initiatives and solutions covering a diversity of use cases,[5] it becomes increasingly important for the field to establish best practices. These practices should ensure, on the one hand, that projects are designed and executed in a responsible, transparent, and privacy-protecting manner, thus being respectful of individuals participating in such research. On the other hand, these practices should also ensure that research projects get meaningful and rigorous answers to the research questions they pose.

The proof-of-concept PORT[6], presented in this article, is an important step towards the development of these best practices. Being among the first to so extensively consider and explicitly implement the notion of local extraction and analysis of DDPs as a generalized practice, its development also points to a set of important considerations that benefit the field more broadly. We briefly discuss some of the main considerations below.

First, *protecting the privacy* of research participants is a core principle that needs to be adhered to throughout the process. In principle, one of the advantages of working with DDPs is that individuals must first download their own data from the appropriate sources and can then decide whether and, if so, how to donate their data for academic research. In other words, scholars do not directly access these packages and instead must rely on the informed consent and the active cooperation of the users.

PORT illustrates how some of these concerns may be by addressed by emphasizing the local extraction and the local processing of the donated data within a participant's machine. This means that, in principle, a researcher may be able to extract relevant measures of behavior (e.g., number of times visiting a news website per period) without the source information (e.g., one's browsing history) ever leaving the participant's computer. PORT not only demonstrates the technical feasibility of this approach but also enables this to run both in desktop *and* mobile devices. It is, however, an open question as to where the greatest risk for breaches in the workflow of data donation procedures lies. The shielding of a user's device from external access, the security settings of the browser used for the extraction script to run in, and the handling of data by researchers after the donations present—as the authors mention themselves—additional privacy risks. PORT addresses an important part in the processing of data donations and at the same time showcases other security challenges that remain throughout the "chain of donation" and that future initiatives need to address.

A second consideration is the *vastness and meaningfulness of the information* researchers can find in a participant's DDP. Data minimization is a second core principle researchers need to adhere to, meaning to only extract and use the data that are necessary for the (ideally pre-registered) research purposes. PORT adheres to this principle insofar as the data extraction follows a predefined script that runs on the participant's device. While data minimization is a core principle that must be part of broader considerations on data donation, it also highlights a dilemma. Researchers may, on the one side of the spectrum, have a clear and narrow definition of the measures that they desire to extract from a DDP (e.g., the *frequency* of usage of news sites within a particular period according to one's browser history and based on a pre-defined list of what the news sites are for a country). On the other end of the spectrum, researchers may need access to non-aggregated level data either because *a priori* classifications may not be sufficient for one's research question (e.g., a *list* of all the domains that a participant has visited, so these domains can be categorized at a later stage), or because the research question itself may not be answerable with aggregate measures in the first place (e.g., inductive analyses of text, or more qualitative approaches).

While some of the issues can be addressed by pre-testing or previous research, misspecifying any of these parameters in a data donation process can

lead to a failure of the project. This is especially problematic, as these research projects are very resource intensive, need a lot of preparation, and for many projects present a one-shot opportunity to get the right data (see van Driel, 2021[7] for an example). As the authors write, striking the balance between the *minimization* of the data while ensuring that the data are meaningful enough to answer the research question, hence, is a second challenge that future initiatives will have to address.

Third, how to ensure *successful and informed participation* in the data donation process is a major challenge. Insights in success rates are sparse, yet some first indication exists: Ohme et al.[8] found 11.6% of a general population sample to donate mobile log data, while van Driel et al.[7] found roughly a fourth of a teenager sample donating their Instagram DDPs. Using the online browsing tracking tool WebHistorian, Wojcieszak et al.[9] gathered 711 donations from the original sample of 3,735 US users. The attrition in data donations endeavors is a crucial concern, not least because it can create sample biases that undermine the quality of data.[8]

PORT highlights the need for a straightforward and user-friendly workflow. Yet, as the authors also indicate, simplifying the user experience is crucial. In addition, next to clear instructions and seamless workflows, trust may be the most important prerequisite for successful donations: participants need not only to trust the researchers but also trust their own technical skills to complete such a process, and importantly, they need to be able to provide *meaningful informed consent* to the usage of their data for academic research. The ability to adequately inform participants in a way that respects their agency in the process may be one of the most important challenges that initiatives on this method have to address. This becomes critical if research is to get meaningful measures not just of a selected tech-savvy few, but rather from a broad and diverse sample of the population—and to do so in a way that ensures that individuals not only *are able* to donate their data, but also *clearly understand and actively consent* to the usage of their data by researchers.

Data donations have the potential to complement existing social science research methods and open exciting opportunities for measures and research projects derived from digital trace data. PORT is an important step in this direction, with its consideration for privacy risks and data minimization. Ultimately, the research community needs several of these initiatives, learning from pitfalls, and the accumulation of experience to arrive at a standard that can make data donations not only an important method for researchers, but especially one that lives up to strict ethical guidelines and that respects and guarantees individual privacy and agency.

DECLARATION OF INTERESTS

Araujo has collaborated with Boeschoten and Oberski in earlier research projects and in a consortium for the further development of data donation infrastructures.

REFERENCES

1. Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., and Oberski, D.L. (2020). Digital trace data collection through data donation. arXiv, 2011.09851 http://arxiv.org/abs/2011.09851.

2. Boeschoten, L., Voorvaart, R., Kaandorp, C., Goorbergh, R.V.D., and De Vos, M. (2021). Automatic de-identification of Data Download Packages. arXiv. 2105.02175. Published online May 4, 2021. https://doi.org/10.48550/arXiv.2105.02175.

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

4. Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., and Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. Nature *595*, 197–204. https://doi.org/10.1038/s41586-021-03666-1.

5. Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., van de Velde, B., de Vreese, C., and Welbers, K. (2021). OSD2F: An Open-Source Data Donation Framework. Published online September 18, 2021. https://doi.org/10.31235/osf.io/xjk6t

6. Boeschoten, L., Mendrik, A., van der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., and Oberski, D.L. (2022). Privacy-preserving local analysis of digital trace data: A proof-of-concept. Patterns *3*, 100444.

7. van Driel, I.I., Giachanou, A., Pouwels, J.L., Boeschoten, L., Beyens, I., and Valkenburg, P.M. (2021). Promises and Pitfalls of Instagram Data Donations (Open Science Framework) [Preprint]. https://doi.org/10.31219/osf.io/krqb9.

8. Ohme, J., Araujo, T., de Vreese, C.H., and Piotrowski, J.T. (2020). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. Mobile Media & Communication *9*, 293–313. https://doi.org/10.1177/2050157920959106.

9. Wojcieszak, M., Menchen-Trevino, E., Goncalves, J.F.F., and Weeks, B. (2021). Avenues to News and Diverse News Exposure Online: Comparing Direct Navigation, Social Media, News Aggregators, Search Queries, and Article Hyperlinks. Int. J. Press/Polit. Published online May 31, 2021. https://doi.org/10.1177/19401612211009160.