Data Sharing Plan

The PIs have a proven track record of releasing research output (scientific and legal publications, memoranda, code, and data) in a timely manner and in ways that are visible, accessible, and understandable by the relevant scientific or legal communities. They will continue this policy of actively sharing information, within the constraints of feasibility and the law.

Types of Data. We anticipate that data produced over the duration of this project will include source code and documentation for software tools as well as processed versions of data collected originally by other sources. During the development of the software, the source code and technical documentation will reside in a public version control repository (e.g., GitHub). Any data generated by the project will be deposited in an archival data repository, such as the Dataverse repository hosted at Harvard University.

Data and Metadata Standards. The data in this project will conform to existing standards where possible, and will conform to clearly documented formats in cases where standards do not exist. The processed versions of data collected originally by other sources and used in the the project's experiments will be converted to R, SPSS or STATA formats. These formats are already fully supported by Dataverse, which performs archival format migration; metadata extraction; and validity checks. Deposit in these formats will also enable on-line analysis; variable-level search; data extraction and re-formatting; and other enhanced access capabilities. Documentation accompanying the experiments or existing data sets will be deposited in PDF/A, or plain-text formats, to ensure long-term accessibility.

Policy for Sharing and Archiving and Preservation of Data. The primary mechanism for archiving and disseminating results will be publications in internationally recognized conferences and journals. We will promptly prepare and submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings from the work proposed here. Preprints of all publications will be deposited in open-access repository (e.g. arXiv) and made available to the public through the website of the Privacy Tools Project. Additionally, source code and documentation for all tools and systems will be made available through GitHub. Intellectual Property rights in general will be managed according to the appropriate university's intellectual property policy or by an open source software license (e.g., MIT). Software contributed to OpenDP will use the permissive and non-viral MIT license. The source code and documentation will be shared in a timely manner.

Archival, long-term access and versioning. GitHub, arXiv and Dataverse currently provide automatic version (revision) control over all deposited data sets and no versions of deposited material are destroyed except where such destruction is legally required.

Reproducibility. To enhance transparency, we endeavor to also store in GitHub everything needed to reproduce our experimental findings. This includes source code of the relevant data analyses, data used in the experiments, and drivers needed to synthesize the results in a graphical or textual form, and the provision of "literate programming" integrations of paper and analysis code (such as Sweave/Pweave, knitr).