STEDI - A Support Tool for Ethical Data Integration

Kavithvajen Kamaraj, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Declan O'Sullivan

After the integration of a few seemingly virtuous datasets, there exists a possibility that the fused dataset might lead to unethical insights and practices. However, the problem is magnified as human data-integrators cannot look at datasets and determine if it might cause an ethics-related issue. This dissertation aims to solve the aforementioned problem by developing a prototype support-tool called STEDI (Support Tool for Ethical Data Integration). Data integrators can use STEDI as a decision support tool to identify ethical issues present in datasets and to also predict the materialisation of ethical problems that could arise from integrating these datasets. The current version of STEDI focuses on working with linked-data datasets. When two or more datasets are loaded into it, STEDI will start by analysing the vocabularies and predicates present in the datasets to understand the kind of data that is present in it. State-of-the-art Natural Language Processing techniques are deployed to understand the data that the predicate entails. The results of the analysis are stored in a specially-designed ethics ontology that can answer a limited amount of questions regarding the ethical views of the datasets. To accurately identify data-integration related ethics issues, the data in the ethics ontology is retrieved to identify four highly-specific ethical issues. Finally, an ethics report is generated containing the results of the ethical analysis. Through evaluation, it is concluded that the proposed approach is viable and can be leveraged in the future to identify ethics-related problems in linked-data datasets.