Innovating in the Public Interest: Courts Data Project

Authors: Rohit Musti, Omkar Bhat, Jack Verrier, Ari Klau **Date:** February 2, 2020

Background

1.1 Context

The original Virginia courts data project - `https://virginiacourtdata.org/` - was created by and is mainatained by Ben Schoenfeld, a computer engineer and civic technology hobbyist. It scrapes records created by the Virginia courts data system and publishes the aggregated results into more easily analyzable representations. This data has been referenced in many news articles to expose bias within the court system and to analyze widespread social problems, like medical debt. Maintaining this project and improving upon it is certainly in the public interest; at the moment, it is supported by Mr. Schoenfeld's private funds and time. We propose coordinating with several groups at the University of Virginia (the Law School's Legal Data Lab, CommPAS Lab of the Batten School & University Library, and the Computer Science department) to host and expand the courts data project.

1.2 Stakeholders

There are two stakeholders for this project: Dr. Michelle Claibourn and Mr. Jon Ashley. Dr. Michelle Claibourn is the Director of Research Data Services at the University of Virginia Library, Co-Director of the CommPAS Lab, and the Director of the Public Interest Data Lab. Jon Ashley is the Head of the Legal Data Lab and Research Librarian. Dr. Claibourn represents the CommPASS Lab, Mr. Ashley represents the Legal Data Lab, and Dr. Jack Davidson along with the students developing the project represent the Computer Science Department.

Project Proposal

2.1 Improvements on the Original Project

We seek to improve on Schoenfeld's original implementation by micro-servicing and containerizing the project. By separating the data scraping, database, and data download points of the project into separate services, we will build a more resilient and easily scalable system. A side benefit is making it easier to understand which datasets are the most popular and identify the most costly steps within the project. If this modular system is designed well, it will also make it smoother for future developers to integrate their own data scraping services. While we haven't finalized the exact implementation details of the project, we anticipate using a relational database and python3. A relational database would make querying and generating CSV files simpler than a non-schema-based database. Python has also become the default language taught within UVA and is also one of the most universal understood programming languages.

2.2 Novel Contributions

We plan on increasing the number of states aggregated for the project. There has already been some work scraping Massachusetts' court data and most other states have open data portals. We anticipate that by creating a national data aggregation tool, even more trends can be discovered. Journalists across the country will also be able to report on either their local trends or pressing national issues using the data. Pushing this project to emcopass is also in line the Public Interest Technology University Network grant that supports this course.

2.3 Risks

The most notable risk by aggregating this data is comprising the privacy of already marginalized groups. We plan on taking steps to remove identifying features from the data. We will also have independent experts review our data processing and data aggregation work to ensure that we are not compromising the rights or privacy of any individuals.