Collecting Publicly Accessible Virginia Court Data into a Searchable and User-Friendly Database

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Jessie Shen
Spring, 2021

Technical Project Team Members
David Alves
Matthew Bacon
Andrew Kim
David Stern

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Jessie Shen

Approved _____ Date _____
Jack Davidson, Department of Computer Science

Collecting Publicly Accessible Virginia Court Data into a Searchable and User-Friendly Database

Many different industries are transitioning to digital organizational solutions to store, organize, and collect sensitive data. Particularly, utilization of digital libraries is on the rise within the legal, medical, and academic areas. Prospective research has been done in determining what makes a system effective and useful.

Elliot & Kling have conducted a study arguing for organizational usability as a useful metric, which is defined as "the mix between a computer system's design and an organization's characteristics such that the system can be effectively integrated into the work practices of members of the organization and is socially accepted by them" (Elliot & Kling, 1996). Our project aims to achieve high organizational usability as this will, ideally, be a useful tool for attorneys or paralegals to conduct their research in one collective database. It should be an additional tool for stakeholders to incorporate into their routine, so ease of use and user experience will be a high priority.

It will be very important to keep these ideas in mind as my team members and I conduct our research on this topic. The professor I am working with, Jack Davidson, has outlined some goals of the project.

These goals are improving the web user interface; developing tools to check data quality; building back-end applications; improved scraping of legal web sites. The technical skills involved in this project include user interface design, Django, Python, Amazon AWS, etc.

I was interested in this project initially as it seemed to have a purpose in serving the public interest in terms of assisting the legal system and making it easy to conduct legal work and research. This could help keep us safe and ensure the law is being properly carried out in the long run. I believe it is important for any technology to ultimately serve the public good, and this project seemed like it would have overall positive impacts.

However, as noted before, this project could also lead to some ethical concerns. Due to the information we are planning to handle being very sensitive, privacy and accuracy of our data scraping algorithm will be of the utmost importance. We will also need to take care to ensure that the tools we use to edit or upload the data are secure, and that nobody will be able to maliciously upload false information to get others in trouble or falsely accuse them of having committed crimes.

In addition, it is important to note that the legal system is not always fair. The goal of this project is to increase fairness and equality, but we risk exacerbating problems in the current system. Hopefully, as long as we keep in mind the importance of validating sensitive information and prevent security breaches or malicious users, we can avoid these issues.

This project has not yet been completed. As I am working on this project for my CS 4980 Capstone Research, I am currently working on this project with a team other students in Spring 2021. For my role in this project, I am responsible for the database aspects, like designing a database schema and writing the code to create the database. I helped to make decisions like what programming language to use (we decided on SQL) and what database service to use. I wrote SQL scripts to create the four databases we will use on the project's virtual machine and server.

Currently, the scraper aspect of the project is nearing completion, with only a few features and bugs left to implement and fix. The scraper was written in Python, and runs a script that opens up the

VA Courts website to methodically go through each district or county court and scrape for case, plaintiff, defendant, hearing, etc. data. Once this scraper is run and the data is collected, we plan to upload this data to a SQL database, and then design a web interface to allow users to search public-facing data.

An important ethical decision we had to discuss was how to handle the differences between public and private data, and if they should be stored in separate databases for privacy and security reasons, etc. The private data includes data on plaintiffs and defendants with full names, addresses, etc. Though this data is all publicly available through the VA Courts website, there are still some concerns about how aggregating names and personally identifiable data into a more searchable, user-friendly form could infringe on the privacy and rights of those individuals. We plan to take measures such as scrambling names and addresses using a UID hash so that they are relatively anonymized.

Though we will not finish this project by the end of the Spring semester, I feel confident that I will have contributed my fair share and made progress towards accomplishing these goals. The project is planning to continue next year.