Interim Report for Web Archiving Award

Filed By

Free Law Project

Berkeley, CA

19 November 2014

Brian Carver – brian@mail.freelawproject.org

I. Summary of Work in Progress and Outcomes Achieved

Columbia awarded Free Law Project two subprojects. The first was to create a "Complete state court opinion harvester," and the second was to create an archive of "federal appellate oral argument audio recordings."

As regards the first subproject, we are pleased to say that it is now complete and has been made publicly available. A blog post announced the completion of this subproject:

http://freelawproject.org/2014/08/12/courtlistener-and-juriscraper-now-support-all-state-courts-of-last-resort/

This blog post and its associated PR were well received by those in the archiving and legal communities and we have since used the system to archive nearly 30,000 opinions from state court websites. All of this content is available via our search interface, our API, or in bulk data files that we recently enhanced, as documented on the personal blog of our lead developer, here:

http://michaeljaylissner.com/posts/2014/11/06/updating-bulk-data-in-courtlistener-more/

Our opinion harvesting code, licensed under the open source BSD license, resides on GitHub here:

https://github.com/freelawproject/juriscraper

and the dramatic expansion of our state coverage enabled by this subproject largely involved additions to the "state" subdirectory here:

https://github.com/freelawproject/juriscraper/tree/master/opinions/united_states/state

When we started work on this subproject we had 87 state scrapers, each focused on collecting the court opinions from a unique state court web page. Many states have multiple courts and operate multiple pages to promulgate various types of opinions from their various courts. Now that we have finished this subproject, we have 154 scrapers focused on state court websites. Thus our state court opinion archiving efforts have nearly doubled through this work.

We are now harvesting content daily from nearly 200 federal and state court websites and gathering precise metadata from each. These websites change from time to time, and maintenance is required to keep the program running smoothly. Since we have been doing similar work for nearly five years, and since we have mature monitoring systems in place, we are confident we will be able to keep this working properly into the future.

One particularly notable achievement occurred with respect to the collection of Alabama court opinions. Alabama is the only state that continues to publish its appellate court opinions *only* through a paid online subscription service. Thus, the general public has for years had essentially no free electronic access to Alabama case law. We were able to talk with those in charge of the system and get a no-cost account from them and permission to archive the material from their closed system. By combining this new archiving effort, enabled through this subproject, with our prior efforts to collect Alabama case law, we now have made public nearly 40,000 opinions from Alabama courts, documents that were previously among the most difficult to obtain via free electronic access.

Additionally, while our proposal focused on achieving coverage of the courts of last resort in each state, usually called the "Supreme Court," we found that we were able in most cases to provide coverage of intermediate courts of appeals within the states as well. We were not certain this would be possible, but are glad to have under-promised and over-delivered.

We believe the harvesting framework we have developed is particularly well-suited for reuse in other contexts, particularly where precise metadata about the harvested resource is essential, but not necessarily easily obtainable from the resource's contents itself. Indeed, our second subproject under this award shows how this framework can be expanded from document resources to audio resources. Furthermore, the type of document or media is largely irrelevant, and customizing this framework for new harvesting contexts gets easier as new examples provide guidance. While we are focused on the products of courts, we can easily imagine expanding our harvesting framework to cover academic journal articles or almost any other type of document or media that one might wish to archive carefully.

The second sub-project under this award is also nearing successful completion and we expect to provide full details on its outcomes in our final report.

II. Any Issues With Project Timeline

No issues have been encountered yet or are expected to affect our ability to complete both subprojects on time. Indeed the first subproject is complete, and the second subproject is nearly complete as of the date of this report.

III. Any Changes in Project Scope or Deliverables

No changes in project scope or deliverables have been made and none are expected. If anything, we have found ways while completing the work to over-deliver additional related deliverables not anticipated when making the proposal.