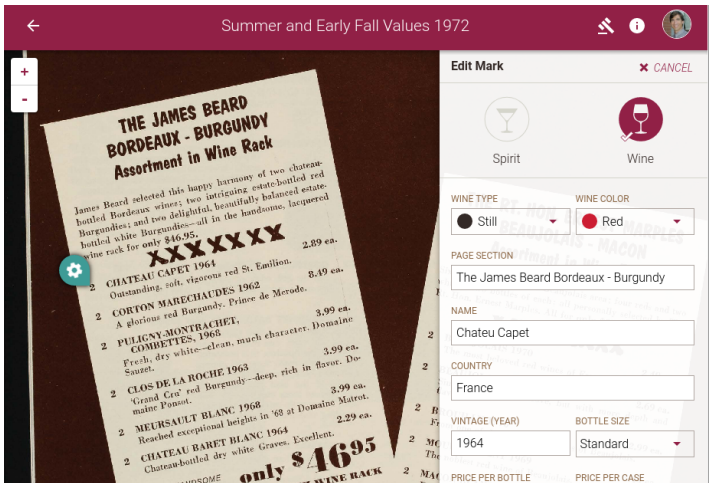
Online Tesseract 4 Extraction

The library is looking to develop an API and application client that accesses library digital materials and provides users the ability to interactively extract text that can be further used in information processing environments. Our development will produce an API that allows interactive text extraction, overlaying those results with the original scanned images. This task can support multiple project or interfaces requiring a combination of OCR with more directed activity.

Background: We have developed crowdsourcing application https://ptv.library.ucdavis.edu, that allows the public to help associate the price with the bottle. Without the benefit of any computer aided textual extraction, users enter wine prices found within the catalogs. This allows our crowdsourcing volunteers to add new wine prices, even when the required information comes from different parts of the page. We have recently explored methods that attempt to extract information from text generated from the page images (above) in a more systematic way. These methods showed some promise but were too specific in their implementation. Interactive integration of OCR results will benefit efforts on three levels. We expect the data extractor to select the sub-section of the page corresponding to the set of wine prices a crowdsourcer is describing. This subsection is scanned in real-time and returned to the user with both the textual information and its location on the page. We can then provide multiple levels of support. First, the user can click on entries in the catalog, and the application can autofill input forms with OCR extracted text. Second, by selecting an common area from the page, simple language processing steps can be used to inform other entries. For example, a language processing step may be able to suggest the wine color, or country of origin from the frequency of words in the selection. Finally, the OCR results can recognize structure, and help the contributor by suggesting complete entries, linking wine names with prices based on their spatial organization on the page.



Requirements: We plan for any project results to be integrated into the libraries application development. We are committed to the maintenance of Javascript / Node applications. In

particular, we develop most of our applications in the polymer web-component framework, In addition, imagery would use a standard Linked Data Platform LDP.

Contact *Quinn Hart* qjhart@ucdavis.edu UC Library Digital Applications Manager