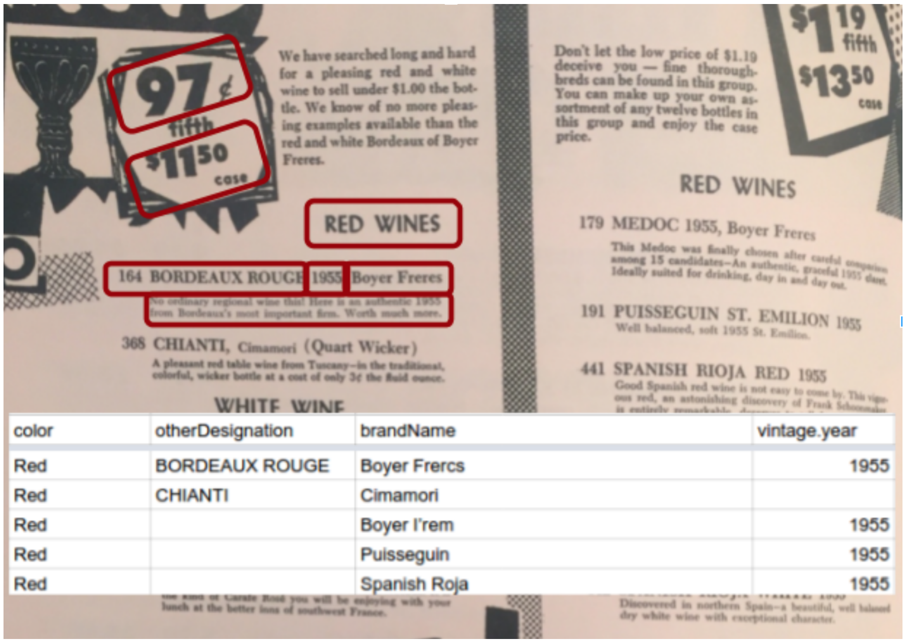
# ECS 193AB Winter/Spring2017

## Template-based Data Extraction



97¢
fifth
$11⁵⁰
case
$1¹⁹
fifth
$13⁵⁰
case

We have searched long and hard for a pleasing red and white wine to sell under $1.00 the bottle. We know of no more pleasing examples available than the red and white Bordeaux of Boyer Freres.

RED WINES

164 BORDEAUX ROUGE 1955 Boyer Freres
No ordinary regional wine this! Here is an authentic 1955 from Bordeaux's most important firm. Worth much more.

368 CHIANTI, Cimamori (Quart Wicker)
A pleasant red table wine from Tuscany—in the traditional, colorful, wicker bottle at a cost of only 3¢ the fluid ounce.

WHITE WINE

lunch at the better inns of southwest France.

Don't let the low price of $1.19 deceive you — fine thoroughbreds can be found in this group. You can make up your own assortment of any twelve bottles in this group and enjoy the case price.

RED WINES

179 MEDOC 1955, Boyer Freres
This Medoc was finally chosen after careful comparison among 15 candidates—An authentic, graceful 1955 claret. Ideally suited for drinking, day in and day out.

191 PUISSEGUIN ST. EMILION 1955
Well balanced, soft 1955 St. Emilion.

441 SPANISH RIOJA RED 1955
Good Spanish red wine is not easy to come by. This vigorous red, an astonishing discovery of Frank Schoonmaker

Discovered in northern Spain—a beautiful, well balanced dry white wine with exceptional character.

| color | otherDesignation | brandName | vintage.year |
|---|---|---|---|
| Red | BORDEAUX ROUGE | Boyer Frercs | 1955 |
| Red | CHIANTI | Cimamori | |
| Red | | Boyer I'rem | 1955 |
| Red | | Puisseguin | 1955 |
| Red | | Spanish Roja | 1955 |

The library is interested in extracting data from scanned data from their historical collections.   These data can thought of as tabular data.  However that is not how they are formatted on the scanned documents, nor do standard OCR reliably capture what structure there is.  Current structured extraction methodologies, like tabula (http://tabula.technology/), do not work well with data that is only semi-structured.   Additionally, most extraction tools are separated from the OCR step.  This can have implications, for example; when you know you are looking only for prices, (eg /\$(\d|[.])+/)

We would like an application or library that would allow the specification of search templates, these templates could be used across multiple scanned documents.  Closely matching templates could be identified for further refinement or curation.  Templates would act on multiple lines, and might associate multiple locations on a document into a single datum.

As ideas, we have considered approximate regular expression searches like TRE (https://en.wikipedia.org/wiki/TRE_(computing)), and implementing the functions as a postgresql extension, using tesseract for OCR, possibly rescanning to better match parts like prices.  We've also considered a javascript library
Quinn Hart <qjhart@ucdavis.edu (mailto:qjhart@ucdavis.edu)> – Digital Applications Manager – UCD Library