Vienna 2019 Abstract Submission

Title
Mining Historic Realia: Automatic Generation of Historic Wine Pricing

I want to submit an abstract for:
Conference Presentation

Corresponding Author
Quinn Hart

E-Mail
qjhart@ucdavis.edu

Affiliation
University of California, Davis

Co-Author/s

Name	E-Mail	Affiliation
Jane Carlen	jacarlen@ucdavis.edu	University of California, Davis
Stanislaw Saganowski	stanislaw.saganowski@pwr.edu.pl	Wroclaw University of Technology
Duncan Temple Lang	dtemplelang@ucdavis.edu	University of Californa, Davis

Keywords
Retail Wine Catalogs, Library Ephemera, Data Modeling

Research Question
Can we leverage computational tools to turn historic wine catalogs into datasets for scholarly exploration?

Methods
Combining text recognition software with pattern recognition, and data cleaning, we investigate strategies for automatic extraction of historic wine prices.

Results
The resulting database will allow researchers to use information from wine catalogs (wine names, prices, etc.) to model data addressing questions in a variety of domains including economics and enology.

Abstract

The University of California, Davis' wine library has been called the greatest in the world. Since 2016, we have been investigating digital applications to showcase and host its wine collections to promote records use and information accessibility. Projects related to these efforts include: "Label This", a crowd-sourced wine label transcription project; new digital collections added to the California Digital Library; and, more recently, "Price the Vintage" [PTV], a crowd-sourced application aimed at extracting historic wine price information from a set of wine catalogs spanning the 50 years from 1938 to 1988. Historical wine pricing data can help in quantifying trends in prices and support their modeling. "Price the Vintage" will create a database of historic prices for various wine types, vintages, countries of origin, and wine-makers. Developing the PTV application has led to an investigation of enhanced solutions for automated derivation. Using the same underlying data, we developed a more sophisticated process to extract additional wine price information.

In this presentation, we will describe the methodologies for both the crowd-sourced application and the automatic extraction of price information. We discuss the pros and cons of each strategy, and how the approaches can be used together. In addition, we will describe some best practices for organizing a digital repository that can be used as a source for these type of projects.

For “Price the Vintage” the UC Davis Library digitized a collection of retail wine and spirit catalogs from the iconic New York City wine shop, Sherry Lehmann, with dates ranging from 1938 through 1988. Among wonderful artwork and articles, the catalogs contain extensive lists of wines and spirits for sale. Uniquely, product numbers in the catalogs often permit the unambiguous tracking of the same wine across multiple catalogs. During digitization, high resolution images were captured for each page in every catalog. Additionally, each page was associated with an initial text rendering by processing the images with the open source optical character recognition [OCR] software, Tesseract 4. These data served as a source for our crowd source application, as well as a new collection in the libraries’ digital special collections repository.

We investigated a number of available crowd-sourcing tools. Our previous “Label This” application was a selection/transcription workflow processing application using the Scribe software, but for PTV we choose an interface inspired in part by the New York Public Library (NYPL)’s What’s on the Menu? Project. The application provides the catalog images as a backdrop and then provides participants with a wine or spirit price data entry form. Crowdsourcing volunteers identify where on the page each particular price is located and interactively fill out the form. For wine, each form includes the following information; wine type; wine color; country of origin; catalog section title; wine name; bottle size; vintage; per bottle price and per case price. Of these parameters: wine type; wine color; country of origin; and bottle size are all limited to a controlled ontology of terms; only the section titles, wine names and prices are entered as free text. For the free text fields, type-ahead suggestions from previous entries are used to speed form-filling times. After considering multiple fields for wines, the simple combination of “section title, wine names” was eventually chosen to reduce the burden of decoding entries for volunteers, allowing them to simply transcribe what they see. Even so, some pages are relatively complex in how information like vintage, country and wine type data are arranged. The crowd source application allows our volunteers to scan pages in context to find this information, resulting in higher quality price data.

However, there are other considerations for crowd sourced data. From our initial crowd-sourcing efforts, we estimate about 17 price entries per page in our catalogs. With 215 catalogs at ~8 pages/item, we can estimate about 50,000 price entries. And while some price information can be entered quite quickly, volunteers with tasks involving many page switches -- e.g., searching for Champagne entries through the set of catalogs -- we find a much lower rate at entering prices. While crowd sourced information has provided us with high quality data for about 1400 entries, we currently are less than 3% complete in capturing the prospective data from this collection.

Many pages in the catalogs have a more standard layout, resembling organized tables of price data. These pages are most amenable to automated price extraction. Originally we investigated general application table extraction tools combined with additional software for table refinement, but generally found that clean-up time exceeded our crowd sourced entry time due to the variance in the source material, made even more burdensome when the refinement step is removed from the context of the original pages.

We have developed a more focused methodology that automates price extraction by dividing the process into three distinct steps. The first step scans pages for individual wine price entries. This step uses OCR results that include positional information for the text discovered on the pages. Pattern recognition techniques then group together wine names, descriptions, prices, dates, and ordering numbers. It’s at this stage where confounding effects like information ordering; changes in columnar organization; multiple lined entries; and OCR transcription confidence are addressed. Distinguishing wine names from surrounding text is perhaps the most challenging aspect of this step. Relevenant words are selected based on heuristics regarding their location and character size in comparison to nearby descriptive text.

The resultant entries are made location independent, grouped, standardized, and initially assigned to fields. These entries then enter the second step where the data is validated. This step includes validating vintage dates and determining individual and case prices. Verifying prices is one of the challenges at this step, as OCR is prone to certain inaccuracies such as confusing 5’s and 9’s. This is addressed though image cropping, pre-processing, and corrections based on context. In addition, this step also looks at local typographic differences to refine its wine name recognition. For example, capitalization and parenthesis often distinguish regions or wine types from other information, but this is not standard from catalog page to page. However, for a group of prices, these conventions can be used to separate components of the wine name.

The step of testing wine names leads into the final step of refinement. We combine our extracted data with other wine source information such as an existing database of tasting notes on thousands of specific wines. We execute fuzzy matches between the catalog names and this database, allowing us to validate proper transcription from the OCR step and adding attributes from our external database to identify fields like wine color, type and region into the original catalog entry.

The crowd-sourced and automatic extraction methods are complimentary for a multitude of reasons. Crowd sourced entries serve as a “truth table”, providing a good test on the accuracy of the automated methods. Not only can we check agreement between the entries, but they also help determine what pages require users to draw information from other sections of the catalog to complete entries, and how the automated method may be refined to capture that information. In addition, automated methods can complete the “simple” pages from the catalogs, while crowd-sourced efforts can be targetted to the more challenging extractions.

Moving forward, we plan to more closely integrate the automated process into the crowd-source application. For example, we plan to update the crowd-source application with additional smaller tasks; for example choosing the best of two wine price markups. Therefore, we can use our volunteers to validate our automated results. We also plan to investigate tools beyond simple pattern matching, such as using machine learning techniques to discover missing fields from a collected wine price.

We will also expand the scope of the historic wine price database. Sherry Lehmann has recently sent an additional hundred catalogs to the UC Davis library with the goal of creating a more complete digital collection of their catalogs. In addition, we have recently digitized another collection from Professor Maynard Amerine: a set of menus and wine-lists from his personal archive. There are more than 1500 menus referencing both private functions and commercial restaurants. We plan to apply the same data protocol to these assets.

Finally, the wine price database will be made available as a tool for modeling. On our UC Davis campus, the library will sponsor a competition for the most innovative use of the data. This project is supported by a generous gift from the Alfred P. Sloan Foundation.

Privacy

- Mining Historic Realia: Automatic Generation of Historic Wine Pricing
- By using this form you agree with the storage and handling of your data by this website.