What IS on the Menu! A Case Study in Data Cleaning and Provenance

Alok K. Shukla,Gitika Jain,Apurva V. Hari

Final Project Report

Abstract—Data Preparation - Cleaning and Wrangling; the first step of any Data Science work-flow is often most time consuming and least enjoyable of all tasks. And that is mainly because these tasks (specially cleaning) largely involve lot of manual steps - mostly repetitive and tedious, which can not be still fully automated; thus tools like Amazon Mechanical Turks [1] and CrowdFlower [2] are quite popular for these reasons. Another issue with whole Data Science work-flow is that of Provenance - lineage of datasets; the experiments are often not reproducible due to lack of proper provenance. The focus of this report is on use of Open Source tools [3] [4] [5] that can help Data Cleaning work-flows be more effective and traceable.

Index Terms—Data Cleaning, Provenance, OpenRefine, YesWorkflow, Workflow, SQL

1 Introduction

This report summarizes our experience with an end-to-end data preparation work-flow; in practice of Data Cleaning and Provenance establishment. We use tools and techniques introduced in CS598 [6]: Theory and Practice of Data Cleaning with a real world dataset [7] and document the whole work-flow along with findings. Tools used include OpenRefine [3], SQLite [5] and YesWorkFlow [4].

2 Dataset overview and initial assessment

2.1 The Dataset

The New York Public Library Rare Book Division holds over 45,000 historical menus. About half of these were collected and curated by Frank E. Buttolph [8] between 1900 and 1921. The menus date from the 1850s to the present and include menus from restaurant, railroad and steamship companies, as well as a range of other organizations.

Beginning in 2011, menus from the NYPL's collection were digitized and transcribed with the help of thousands of volunteers. Through the NYPL's What's on the Menu? [7] project, volunteers looked at digitized copies of the menus and typed in the many pieces of information included on each one, such as restaurant names, locations, dishes, prices, and dates.

- *Alok K. Shukla,*
  E-mail: alokks2@illinois.edu,
- *Gitika Jain,*
  E-mail: gitikaj2@illinois.edu,
- *Apurva V. Hari,*
  E-mail: vhari2@illinois.edu,
  University of Illinois at Urbana–Champaign.

Manuscript received on Aug 4, 2017.

The What's on the Menu? [7] project makes all the data from its crowdsourced transcriptions available via bulk downloads and via an application programming interface (API). The current data set includes around 400,000 data points from the transcription project and the library's metadata on the over 17,000 menus digitized so far. The dataset is updated twice monthly, the dataset that we used was released on Jun 17, 2017. It contains 4 files : Menu.csv, MenuItem.csv, MenuPage.csv and Dish.csv .

Menu
: The core element of the dataset. Each Menu has a unique identifier and associated data, including data on the venue and/or event that the menu was created for; the location that the menu was used; the currency in use on the menu; and various other fields. Each menu is associated with some number of MenuPage values.

MenuPage
: Each MenuPage refers to the Menu it comes from, via the menu_id variable (corresponding to Menu:id). Each MenuPage also has a unique identifier of its own. Associated MenuPage data includes the page number of this MenuPage, an identifier for the scanned image of the page, and the dimensions of the page. Each MenuPage is associated with some number of MenuItem values.

MenuItem
: Each MenuItem refers to both the MenuPage it is found on – via the menu_page_id variable – and the Dish that it represents – via the dish_id variable. Each MenuItem also has a unique identifier of its own. Other associated data includes the price of the item and the dates

when the item was created or modified in the database.

Dish

A Dish is a broad category that covers some number of MenuItems. Each dish has a unique id, to which it is referred by its affiliated MenuItems. Each dish also has a name, a description, a number of menus it appears on, and both date and price ranges.

2.2 Use-cases discussion

2.2.1 Is data clean enough already?

The data is really messy for any actual practical use-case other than getting an overall sense of number of menus and dishes from a particular time period - since the date columns even though not perfectly clean can be used to get an aggregate sum. Few more -

- Analyzing the dish dataset can provide the information about the popular dishes based on how many times the dish has been appeared in a menu, when was the first/last time the dish appeared in the menu.
- Menu Page can be used to gather the information about the menu structure like height, width in previous years.
- Menu Page and Menu item together can be used to get the information about a particular menu item when and where it appeared on a menu.

2.2.2 Potential use-cases of cleaned Data

Once properly cleaned, we can discuss the possibility of working on use-cases like -

- Can be used to predict the food preferences over the time based on the event, years and location.
- How the Popularity of a dish changed over the time.
- The structure of the Menu.
- How has the median price of restaurant dishes changed over time?Are there particular types of dishes (alcoholic beverages, seafood, breakfast food) whose price changes have been greater than or less than the average change over time?
- Can we predict anything about a dish's price based on its name or description?
- There's been some work on how the words used in advertisements for potato chips are reflective of their price; is that also true of the words used in the name of the food?
- Are, for example, French or Italian words more likely to predict a more expensive dish?

3 Data Cleaning with OpenRefine

3.1 Menu

*3.1.1 **sponsor***

- Trim leading and trailing white spaces.
- Collapse consecutive white spaces.
- Convert column values to upper case
- Remove the special characters.
- Replace Semicolon (;) with a space and then trim leading and trailing white spaces and collapse consecutive white spaces.
- Make a facet and perform the cluster operation using the **key-collision** method and **fingerprint** function. Merge the relevant clusters.
- Repeat last step with **n-gramfingerprint, metaphone3, cologne-phonetic** methods.
- Make a facet and perform the cluster operation using the **nearest neighbor method** and **levestein distance** function . Merge the relevant clusters.
- Make a facet and perform the cluster operation using the nearest neighbor method and **PPM distance** function. Merge the relevant clusters.

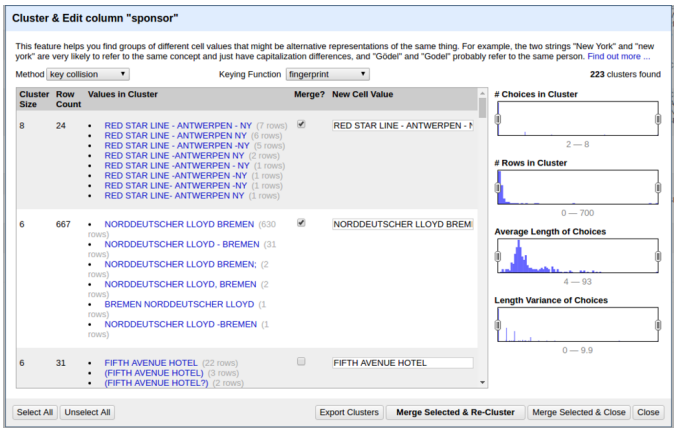

Figure 1. Clustering "sponsor"

Repeat same steps for **event, venue, place, occasion** and **location**.

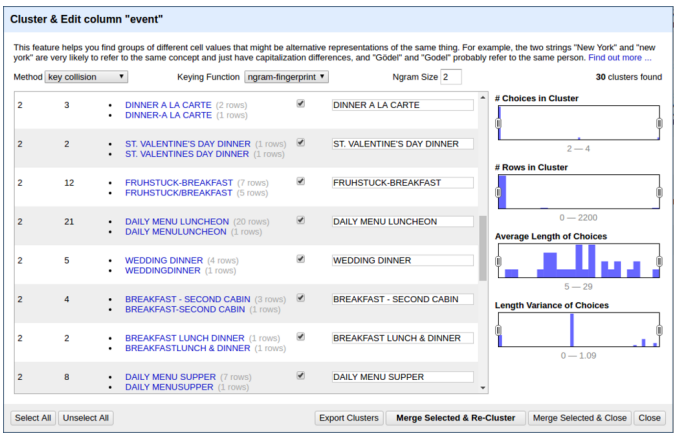

Figure 2. Clustering "event"

*3.1.2 **physical_description***

- Split the column values using semicolon (;) as separator, we get 7 columns as a result.
- Rename the first column as "physical_description_type".

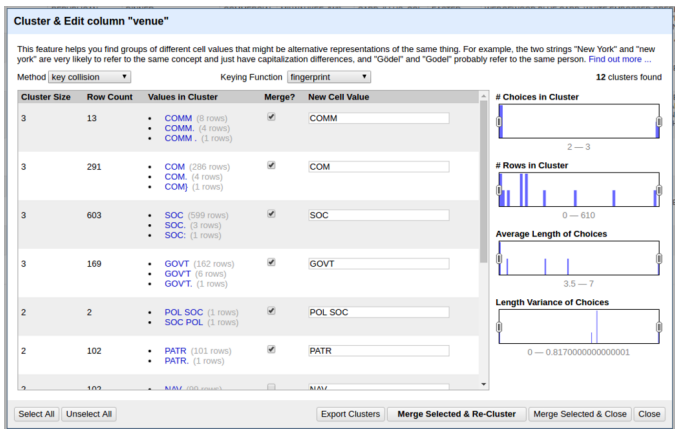

Figure 3. Clustering ”venue”

- Using the GREL function, join only the separated columns - ”physical_description 2”, ”physical_description 3”, and ”physical_description 4” into one column, separate the value using a dash (-) character and name the column as ”physical_description_additional”. Leave the respective value as blank space if the column is null or blank.
- In case, if the ”physical_description 2” column is empty, then make ”physical_description_additional” empty as well.

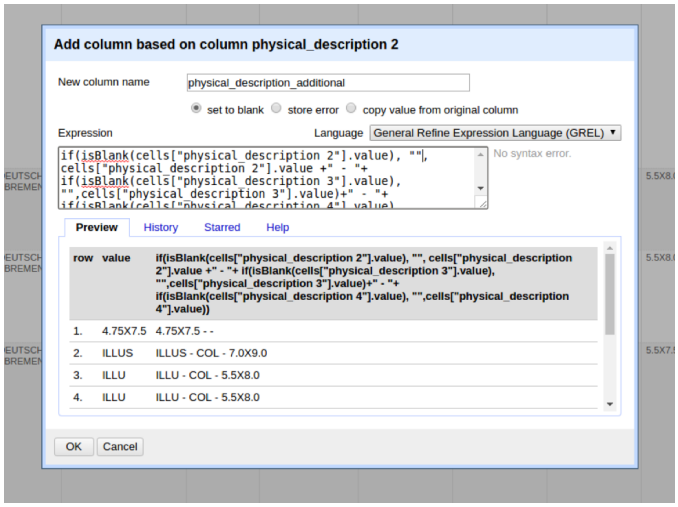

Figure 4. Cleaning ”physical_description”

3.1.3 *date*

- Convert the ”date” column values to date format ”YYYY-MM-dd”. Also, get rid of outliers where the year is “less than 1851 and more than 2012”.
- Replace the outliers with empty characters.

3.1.4 *Left as-is*

id , name, notes, call_number, keywords, language, status,page_count, dish_count.

3.2 MenuPage

No column cleaned with OpenRefine.

3.3 MenuItem

The **created_at** and **updated_at** columns were cleaned as **date** column of **Menu** file.

3.4 Dish

The **name** column was cleaned as **sponsor** column of **Menu** file; and **first_appeared**, **last_appeared** as **date** column of **Menu** file.

4 Developing a Relational Schema with SQL

TO-DO Apurva

Duis rhoncus velit nec est condimentum feugiat. Donec aliquam augue nec gravida lobortis. Nunc arcu mi, pretium quis dolor id, iaculis euismod ligula. Donec tincidunt gravida lacus eget lacinia.

5 Creating a Work Flow Model with YesWorkflow

We used the YesWorkflow on-line editor [4] to create the workflow graph for whole process.

5.0.1 *The Complete Work-flow Graph*

The complete graph generated by YesWorkflow.

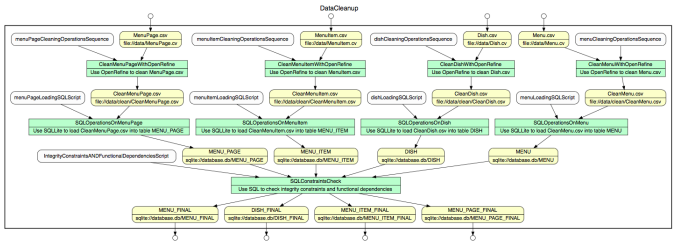

Figure 5. The Work Flow Graph

5.0.2 *Data-flow Graph*

Without the operations - Data Lineage.

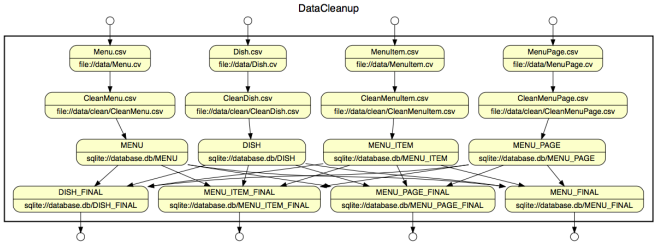

Figure 6. The Data Flow Graph

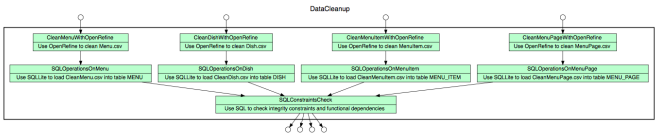

Figure 7. The Operations Graph

5.0.3 *Operations Graph*

Without the data - only operations.

6 Alternate Approaches

Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit amet dui dui. Duis rhoncus velit nec est condimentum feugiat. Donec aliquam augue nec gravida lobortis. Nunc arcu mi, pretium quis dolor id, iaculis euismod ligula. Donec tincidunt gravida lacus eget lacinia.

Write here Some text. This is a bibliographic citation [**?**]. Duis rhoncus velit nec est condimentum feugiat. Donec aliquam augue nec gravida lobortis. Nunc arcu mi, pretium quis dolor id, iaculis euismod ligula. Donec tincidunt gravida lacus eget lacinia.

Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit amet dui dui. Duis rhoncus velit nec est condimentum feugiat. Donec aliquam augue nec gravida lobortis. Nunc arcu mi, pretium quis dolor id, iaculis euismod ligula. Donec tincidunt gravida lacus eget lacinia.

7 Conclusion

The conclusions. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit amet dui dui.

Duis rhoncus velit nec est condimentum feugiat. Donec aliquam augue nec gravida lobortis. Nunc arcu mi, pretium quis dolor id, iaculis euismod ligula. Donec tincidunt gravida lacus eget lacinia. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Acknowledgments

The authors would like to thank...Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit amet dui dui. Duis rhoncus velit nec est condimentum feugiat. Donec aliquam augue nec gravida lobortis. Nunc arcu mi, pretium quis dolor id, iaculis euismod ligula. Donec tincidunt gravida lacus eget lacinia.

References

[1] "Amazon mechanical turk," https://www.mturk.com/mturk/welcome.
[2] "Crowdflower," https://www.crowdflower.com/.
[3] "Openrefine: A free, open source, powerful tool for working with messy data," http://openrefine.org/.
[4] B. L. et al, "Yesworkflow," https://github.com/yesworkflow-org/.
[5] "Sqlite," https://www.sqlite.org/.
[6] B. Ludaescher, "Cs598 - special topics," https://cs.illinois.edu/courses/profile/CS598, Summer 2017.
[7] N. Y. P. Library, "What's on the menu?" http://menus.nypl.org/, 2011.
[8] "Frank e. buttolph," https://en.wikipedia.org/wiki/Frank_E._Buttolph.

Alok K. Shukla Here I am. I am pursuing my Engineering studies at **IST!** (**IST!**). Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit amet dui dui. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Gitika Jain Here I am. I am pursuing my Engineering studies at **IST!**. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit amet dui dui.Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Apurva V. hari Here I am. I am pursuing my Engineering studies at **IST!**. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras sed sapien quam. Sed dapibus est id enim facilisis, at posuere turpis adipiscing. Quisque sit amet dui dui.Lorem ipsum dolor sit amet, consectetur adipiscing elit.

APPENDIX