CS 513 Theory & Practice of Data Cleaning

Final Project Report: What’s On The Menu?

Artsiom Strok  
NetID: astrok2   
[astrok2@illinois.edu](mailto:astrok2@illinois.edu)  
Ramin Melikov  
NetID: melikov2  
[melikov2@illinois.edu](mailto:melikov2@illinois.edu)Jonah Willis  
NetID: jonahww2  
[jonahww2@illinois.edu](mailto:jonahww2@illinois.edu)

*Abstract*— Data cleansing is one of the most important and most time-consuming steps in data science projects. The final project aims to use various tools and techniques covered in this course, together in an end-to-end data cleaning workflow. All knowledge obtained in this course will be applied for New York Public Library’s crowd-sourced historical menus dataset.

Keywords—NYPL, menus, data cleansing, OpenRefine, YesWorkflow, SQL, provenance.

# Overview and initial assessment of the dataset

The New York Public Library is digitizing and transcribing its collection of historical menus. The collection includes about 45,000 menus from the 1840s to the present, and the goal of the digitization project is to transcribe each page of each menu, creating an enormous database of dishes, prices, locations, and so on. As of early July, 2020, the transcribed database contains 1,334,431 dishes from 17,545 menus.

This dataset is split into four files to minimize the amount of redundant information contained in each (and thus, the size of each file). The four data files are *Menu, MenuPage, MenuItem*, and *Dish*. These four files are described briefly here, and in detail in their individual file descriptions below.

## 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.

* id – The unique identifier of the menu
* name – The name of the restaurant
* sponsor – Who sponsored the meal (organizations, people, name of restaurant)
* event – The category (e.g. lunch, annual dinner)
* venue – The type of place (e.g. commercial, social, professional)
* place – Where the meal took place (often a geographic location)
* physical\_description – The dimension and material description of the menu
* occasion – The occasion of the meal (holidays, anniversaries, daily)
* notes – The notes by librarians about the original material
* call\_number – The call number of the menu
* keywords – The keywords of the menu
* language – The language of the menu
* date – The date of the menu
* location – The organization or business who produced the menu
* location\_type – The type of the location
* currency – The system of money the menu uses (dollars, etc.)
* currency\_symbol – The symbol for the currency ($, etc.)
* status – The completeness of the menu transcription (transcribed, under review, etc.)
* page\_count – How many pages the menu has
* dish\_count – How many dishes the menu has

Each menu is associated with some number of MenuPage values.

The first inspection of the data shows us that this file has 17545 entries and 20 columns.

Three columns ‘*keywords*’, ‘*language’*, ‘*location*\_type’ do not have any values and can be deleted.

The ‘id’ column has all unique numeric values. Thus we can assume no issues with this column.

The ‘*name*’ column has only 3197 non-empty values. There also placeholders for missing value, e.g., *‘[Restaurant name and/or location not given]*’ or *‘[Not given]*’. There are a lot of names that exactly the same but due to extra spaces, punctuations, different order of words, typos, diacritical mars they don’t match exactly.

The ‘*sponsor*’ column has 15984 non-empty values, and these values have similar issues as ‘*name*’ column. Also, some of the values are just question marks.

The ‘*event*’ column has 8154 non-empty values. The values for this column can be grouped into different buckets such as ‘*breakfast*’, ‘*lunch*’, ‘*dinner*’ etc. Also, some of these values are written in different language e.g., French or German, and depends on the use case can be grouped together. The values such as ‘*107th, 108th ... anniversary dinner*’ can be grouped together as just ‘*anniversary dinner’*. Each value can have multiple categories e.g., ‘*lunch and dinner’*, which also can be post-processed based on the use case.

The ‘*venue*’ column has 8119 non-empty values. The values in this column have the most of common issues, including question marks, extra punctuations, etc., and new unique issues with abbreviations e.g., ‘*SOC*’ and ‘*SOCIAL*’, ‘*COM*’ and ‘*COMMERCIAL*’. In addition, this column can also have multiple categories within one value. The ‘*place*’ column has 8123 non-empty values. And again, besides common issues, this column has an issue with partial values. The value can represent just the name of the place or place and city or address line, city and state, etc.

The ‘*physical\_description*’ column 14763 non-empty values. There as some *‘#N/A*’ values. Each value in this column has multiple sub-values such as type of menu e.g. ‘*booklet*’, ‘*card*’, ‘*folder*’ and physical dimensions of the menu e.g. ‘*5.75 X 7.25*’, ‘*5 X 8*’ and some unique features of the menu e.g. with or without illustration, regular or column layout, folded or open. And this column can have multiple variations of such properties within one value.

The ‘*occasion*’ column has 3791 non-empty values. The values of this column also can be grouped into multiple buckets.

The ‘*notes*’ column has 10613 non-empty values. The values in this column mostly represented by paragraphs of free text, mostly unstructured. Depends on the use case, additional features can be derived for this column.

The ‘*call\_number*’ column has 15983 non-empty values. The majority of values in this column are numeric with some OCR-like issue e.g. we see ‘*o*’ instead of ‘*0*’, or ‘*l*’ instead ‘*1*’. Some of them have postfixes such as ‘*item*’, ‘*\_wotm*’, ‘*copy*’. And some of them starting from the word and continuing with a number, e.g. ‘*Zander 645*’, ‘*Soete 162*’, ‘*Baratta 35*’.

The ‘*date*’ column has 16959 non-empty values. And only three values where there are some issues with the year and can be easily detected using timeline facet from OpenRefine. The ‘location’ column does not have empty values. However, there are values such as question mark. The issues are similar to the issues with ‘*name*’ or ‘sponsor’ columns.

The columns ‘*currency*’ and ‘*currency\_symbol*’ both have 6456 non-empty values, and they look good. Some preprocessing can be done for cents because it can be cents of different currency.

The ‘*status*’ column has all values available and does not have any issues.

The ‘*page\_count*’ and ‘*dish\_count*’ columns also have all values available. There are some extreme values that need to be analyzed.

## 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.

* id – The unique identifier of the menu page
* menu\_id – The unique identifier of the menu, corresponds to Menu id
* page\_number – The number representing sequence of page in the menu
* image\_id – The unique identifier of the page image
* full\_height – The height of the page image in pixels
* full\_width – The width of the page image in pixels
* uuid – The universally unique identifier for the highest resolution version of the image

Each MenuPage is associated with some number of MenuItem values.

The first inspection of the data shows us that this file has 66937 entries and seven columns. The *id* and *menu\_id* columns seem clean on first inspection with no missing entries and a relatively uniform distribution.

The *page\_number*, *full\_height*, and *full\_width* columns all have missing entries (1202, 329, and 329, respectively) but seem to be otherwise clean. Both *full\_height* and *full\_width* are missing entries in the exact same rows.

The *image\_id* column presents the most issues. The values in this column are using three different formats. About half of the entries are using 7-digit numeric IDs, another half are using 10-digit numeric IDs, and a few (23) of the values are using alpha-numeric IDs.

Finally, the *uuid* column was almost entirely clean, only one entry needed to be updated to use lower-case letters. It is worth noting that some uuids are duplicated.

## MenuItem

Each MenuItem refers to both the MenuPage it is found on -- via the menupageid 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.

* id – The unique identifier of the menu item
* menu\_page\_id – The unique identifier of the page the menu item is on, corresponds to MenuPage id
* price – The first price of menu item
* high\_price – If the item has more than on price on a single menu, the highest price. If there are more than two values for price, the web application instructs volunteers to enter the lowest and highest prices rather than all values.
* dish\_id – The unique identifier of the dish, corresponds to Dish id
* created\_at – The date/time of the first transcription
* updated\_at – The date/time of the last edit to the value
* xpos – The horizontal coordinate on the page for the upper left point where menu item is on the page
* ypos – The vertical coordinate on the page for the upper left point where the menu item is on the page

The first inspection of the data shows us that this file has 1332726 entries and nine columns. The *id*, *menu\_page\_id*, *dish\_id*, *created\_at*, *updated\_at, xpos,* and *ypos* columns seem clean on first inspection.

The *price* column 445916 blank rows. It is also worth noting that there are 130 rows with extremely high (over $10000) prices.

The *high\_price* column has 1240821 blank rows, which means that the vast majority of the rows are blank. It may be worth excluding this column.

## 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.

* id – The unique identifier of the dish
* name – The name of dish
* description – The description of the dish
* menus\_appeared – The total count of menus on which dish with this id appears
* times\_appeared – The total count of appearances of the dish with this id across all menus
* first\_appeared – The earliest year of a menu on which a dish with this id appears
* last\_appeared – The latest year of a menu on which a dish with this id appears
* lowest\_price – The lowest price associated with a dish with a given id
* highest\_price – The highest price associated with a dish with a given id

<http://menus.nypl.org/about>

<https://www.kaggle.com/nypl/whats-on-the-menu>

The ‘*Dish.csv*’ file has data about dishes that have appeared or didn’t appear in menus since 1851. In particular, it contains data about the dish’s name, its description, how many times and in how many menus it appeared, when it appeared first and last, and its lowest and highest price. This dataset contains 423,397 observations of 9 variables. The variables are as follows:

* id – The unique identifier
* name – The name of the dish
* description – The description of the dish
* menus\_appeared – The number of menus in which this dish appeared
* times\_appeared – The number of times this dish appeared
* first\_appeared – The year when this dish appeared first
* last\_appeared – The year when this dish appeared last
* lowest\_price – The lowest known price for this dish
* highest\_price – The highest known price for this dish

Upon initial analysis, some data quality issues have been discovered. First, the ‘description’ column has no values in it at all. And at first thought one would think that perhaps it should be removed. However, when we analyzed the ‘name’ column, we have discovered that there are 9,125 rows where the name column most likely contains the description of the dish because the length of the text is over 100 characters and most of the names are under 100 characters. This is an issue because if we move the text of these rows from the ‘name’ column into the ‘description’ column, then how do we fill the ‘name’ column? One way to do it would be to go through all of the descriptions and then try to deduce a name for the dish from the description but this would require quite a bit of work. One possible solution would be to move the text into the description column and fill the ‘name’ column with the ‘*Unknown*’ value. However, when we think of it in parts of the dataset, 9,125 represents just 2% of the data. That is, only 2% of the data has a description. In that case, does a description have value if it only 2% of the data has it? Maybe this could be normalized further and perhaps there is a substructure there, but it is hard to tell. Furthermore, there are 48,311 duplicates in the ‘*name*’ column. Should they be excluded and if so, what about the statistics they contain?

The ‘*times\_appeared*’ column has some negative values, suggesting that some dishes appeared ‘*-2*’ times in some menus while others ‘*-6*’, etc. There are also some ‘*0*’s in there as well. One recommendation would be to lump all of the values less than or equal to 0 into the ‘*0*’ group. However, there is another problem with this column. For many of the ‘*0*’ values, there are values of ‘*0*’, ‘*1*’, ‘*2*’, and ‘*3*’ in the ‘*menus\_appeared*’ column. This doesn’t make sense. How could a dish appear 0 times but appear 1 time in a menu? Maybe these 2 columns could be merged into 1?

There are also issues between the ‘*first\_appeared*’ and ‘*last\_appeared*’ columns. First, in the ‘*first\_appeared*’ column we have some values that don’t fit in. They are ‘*0*’, ‘*1*’, and ‘*2928*’. All other values in this column fall in the range between 1851 and 2012. There is a same problem in the ‘*last\_appeared*’ column. Furthermore, upon some testing of the values, we’ve discovered that some values in the ‘*first\_appeared*’ column are greater than those in the ‘*last\_appeared*’ column. Granted, there aren’t that many that comprise this violation.

And finally, there are some issues in the ‘*lowest\_price*’ and ‘*highest\_price*’ columns. Mainly, we have some quite a bit of blank rows there for each column. However, we’ve test to see if there are any violations in the data like ‘*lowest\_price*’ greater than ‘*highest\_price*’ or if there are values in another while there are blanks in one but there weren’t such violations here.

Some use cases for this data would be in the space of restaurant entrepreneurs. Before a restaurant offers a dish, it could look at this dataset to see if a similar dish has been in the menus and how popular it is, as well as the price at which it’s been offered throughout the years.

Other use cases would be in the space of journalists or other researchers who are doing research on some dishes. It would be interesting to see if there are any forgotten dishes that perhaps could be revived.

# Data cleaning with OpenRefine

We use OpenRefine version 3.1 to clean the dataset. Each file was cleaned separately.

## Menu file

The ‘*id*’ column was converted to the number using common transformation ‘*To Number*’. See Figure 1 All 17545 were converted without any issues as expected.

A screenshot of a cell phone

Description automatically generated

Figure 1 Common transform for number

Common transformations such as ‘*Trim leading and trailing whitespace’*, ‘*Collapse consecutive whitespace*’, ‘*To titlecase*’ and in some cases ‘*To uppercase*’ were applied for each text column. The same transformations were applied repeatedly when other transformations may cause the issues fixed before.

A screenshot of a cell phone

Description automatically generated

Figure 2 Common transforms for text

In the next step, we applied the text filter ‘*[^\w\s]*’ with ‘*regular expression*’ checkbox checked for each column to identify values containing the non-word or special characters ‘*"&[](){}?%#!/*’. Later all special characters were removed by GREL expression ‘*value.replace(/["&(){}?%#!\/\[\]]/, "")*’.

A screenshot of a cell phone

Description automatically generated

Figure 3 Text filter to keep values with spacial characters

A screenshot of a cell phone

Description automatically generated

Figure 4 Custom text transform using GREL to remove special characters

The rest of the punctuations were replaced by space. Common transformations to remove leading and trailing spaces, and collapse consecutive spaces, were applied.

The values *‘[Restaurant name and/or location not given]*’ or *‘[Not given]*’ were replaced by empty value.

All diacritic characters were mapped to ASCII characters using standard mapping e.g., ‘*á*’ mapped to ‘*a*’, ‘*é*’ to ‘*e*’.

The most of transformations were done using the text facet and cluster feature of OpenRefine. The *‘key collision*’ method with all four key functions ‘*fingerprint*’, *‘ngram fingerprint*’, ‘*metaphone3*’, *‘cologne-phonetic*’, and *‘nearest neighbor*’ method with ‘*levenshtein*’ and ‘*PMM*’ distance functions with default parameters were used to find and group values.

A screenshot of a social media post

Description automatically generated

Figure 6 Clustering of text values

The ‘*date’* column was converted to date format. Two outliers were identified using timeline facet and corrected manually by looking at original images.

A screenshot of a cell phone

Description automatically generated

Figure 7 Common trasform for date

A screenshot of a cell phone

Description automatically generated

Figure 8 Timeline facet to detect outliers

|  |  |
| --- | --- |
| Column | Cells Affected |
| name | 1481 |
| sponsor | 11375 |
| event | 1448 |
| venue | 2551 |
| place | ? |
| physical\_description | ? |
| occasion | 2616 |
| notes | 10613 |
| call\_number | 15981 |
| keywords | 0 |
| language | 0 |
| date | 16959 |
| location | ? |
| location\_type | 0 |
| currency | 0 |
| currency\_symbol | 0 |
| status | 0 |
| page\_count | 0 |
| dish\_count | 0 |

For more details see ‘*Open\_Refine\_History-Menu.json’ file.*

# Data cleaning with other tools

If you find that certain data cleaning steps are not well suited for OpenRefine (e.g. due to scalability or other issues), consider using an alternative, more suitable solution, e.g., Python, R, or other tools such as Trifacta Data Wrangler, Tableau, etc. Document your choice and answer the same questions as in Step 2.

# Developing a relational schema

Develop a relational schema for your dataset. What logical integrity constraints (ICs) can you identify? Load the data into a SQLite database with your target schema. Use SQL queries to profile the dataset and to check the ICs that you have identified! You can also use other query languages such as Datalog to profile the dataset and check the ICs, but you should not use a procedural language such as Python, R, etc.

A screenshot of a computer

Description automatically generated

Figure 9 ER Database Schema

# Creating a workflow model

Create a workflow model of your overall data cleaning workflow: What are the key inputs and outputs of your workflow? What are the dependencies? Note: Here you may want to model the various steps you have executed with OpenRefine as parts of the workflow. This way, the workflow model more clearly describes what actually happened to what parts of the data. Create a visual representation of your overall workflow using YesWorkflow or other diagramming tools. Supplementary material to help with YesWorkflow will be posted on Piazza. Also create a visual representation of your OpenRefine workflow using OR2YWTool (https://pypi.org/project/or2ywtool) or other appropriate tools. The OR2YWTool provides an auto-parsing method from Openrefine Operation History JSON file to YesWorkflow model (developed by Lan Li and Nikolaus Nova Parulian). Please include both overall workflow and OpenRefine workflow in your project report.

# Developing provenance

Develop provenance queries (in Datalog / DLV) that show on which inputs and intermediate data and steps the outputs of your workflow depend (cf. Provenance Assignment).

# Contribution

Develop provenance queries (in Datalog / DLV) that show on which inputs and intermediate data and steps the outputs of your workflow depend (cf. Provenance Assignment).