**Data Clorax  
CS513 Data Cleansing**

Final Project Report

July 5, 2020

Asad Bin Imtiaz, Muhammad Rafay

# Introduction

This report summarizes the methods and tools, as well as the analysis and findings carried out for data wrangling, standardization and provenance workflow for the [1] *The New York Public Library* (NYPL)*, What's on the menu?* Dataset*.* The dataset can be download from the NYPL GitHub website [2]. The analysis and findings are part of Final Project for CS513: Theory and Practice of Data Cleaning Course from the university of Illinois, wherein the NYPL dataset [1] was used to create an end-to-end data wrangling and provenance workflow, together with data landscape analysis and findings, using machine learning and data cleansing techniques learned in the class. The goal of this project was to several of opened-sources Machine learning, Data Wrangling and Data Provenance tools to come up with Data Cleaning Workflow which effectively cleans the selected dataset to high quality standard with all the lineage and audit tracing available.

## Tools and libraries

Following tools were used in this report:

* Python 3 with Jupyter notebook
* OpenRefine data cleaning tool [3]
* SQLLite [4] with DB-Visualizer Pro 9.2 [5]
* Yes Workflow [TBA]
* [TBA] more tools here

## Dataset

The New York Public library (NYPL) maintains a large collection of Menus (~45K) in their 'What's on the Menu' [1] dataset, which is openly available to download [2]. This include dish-by-dish menus from variety of businesses from as early as 1850, and are used by historians, nutritionists and researchers around the globe to understand the patterns and to answer specific questions. The data is collected by taking photographs of menus aver several years by volunteers and was digitized in the dataset form in NYPL Digital Gallery [1].

As with all the crowd-sourced gathered data, there are several gaps and inconsistencies in the data, as well as areas with potential for improvement in terms of the data formats, linking & lineage and its schema.

The entire dataset consists of four character-delimited files:

1. **Dish.csv**

This file contains dish names listed on the menu along with their respective pricing and chronology information. Each dish has an identifier which uniquely identifies it.

1. **MenuItem.csv**

This file contains menu items which link a menu page entity with dish entities. Each record is identified by a unique identifier and carries other information such as associated dish price and x/y position on image of menu page.

1. **MenuPage.csv**

This file contrains menu page records. Each item is identified by a unique identifies and links a menu item with a menu. Additional information such as page photo image number and page dimensions also appear here

1. **Menu.csv**

This file contains all individual menus, each associated with a unique id. Associated data includes the occasion, venue and event information and chronological information such as created and updated dates and times.

## Approach

The project work was divided into multiple tasks. Below is the task breakdown:

* Overview and initial assessment of the dataset.
* Data cleaning with OpenRefine
* Data cleaning with other tools
* Developing a relational schema
* Creating a workflow model
* Developing provenance

Each of the sub task is discussed as separate chapter in the following.

# Overview and initial assessment of the dataset

Issues in Data:

**Dish:**

Cleansing in names required.

1> Same name in Upper and Lower Cases

e.g.

18958 Lobster salad

463235 Lobster Salad

2> Names appear winthin and without quotes

e.g.

''" " saute au madere'

3> Extra Trailing/Leading/between spaces in names.

e.g.

' " hashed in cream'

4> Character set cleansing

e.g.

'SoufflÃ© d'Volaille Ã  l'Artillerie'

5> Cleanse invalid names:

e.g.

'" " 5 Star'

Cleansing in description required.

1> All nulls. Filed to be removed

Cleansing in menus appeared.

1> the counts are not correct in several cases

e.g.

dish id 19 appeared in 16 menus instead on 15

Some issue may be solved with other cleasings in all tables.

Cleansing in first\_appeared.

1> contains year of first appearance. Value is 0 (not known) in many cases

2> It has to be validate (or may be overwritten) with minimum created\_at value in Menu\_Items assiciated with dish

Cleansing in last\_appeared.

sa me as first appeared

Cleansing in lowest/highest price.

1> to be validated/overwritten by associated menuitems

Good:

No PK Uniqueness voilation

**Menu Item:**

Foreign key voilation, 3 cases in raw data with dish

No foreign key voilation with menupage

validation of High price. Must not be less than price

validation of CTEATED/UPDATED date. Created Must not be later than updates

standardazition of timestamps with zones

understand semantic of Xpos,ypos and standardize.

**Menu Page**

UUID is always null

Missing, null values in height,width fort a page id

updated at is a UUID and not date. Maybe an invalid column mapping.

**Menu**

Many names are blank

Standardize sponsor, event (Title case etc), characterset etc.

Missing values in venue. Standardize

standardize place

understand physical desc. denormaloze

understand iccasion

standardize call numver

understand keywords, langiage. standardize if necessary

normalize locations, types

normaloze (ISO Standard) currency

understand status

validate page count with menu page

validate dish count

validate Fk,PK constiaints

# Bibliography

|  |  |
| --- | --- |
| [1] | NYPL Labs, "What's on the menu Dataset," 16 June 2020. [Online]. Available: http://menus.nypl.org/data. |