**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 downloaded 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 use several open-source 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]. The dataset consists of CSV files with entities such as dish-by-dish menus from a 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 over 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 goal of this project is to identify the issues and fix them, keeping the provenance and transformation lineage to understand the cleansing workflow and to later reproduce the cleaned dataset on newer dataset versions.

The initial assessment of data quality and respective issues are presented in chapter 2 in detail.

## 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

In the following sub section, the structure and content on the dataset is inspected before starting with the data wrangling and provenance workflow, to get familiarity with data schema and a feel for apparent data quality issues present in the data. There may be more issues in data which would be discussed in subsequent chapters with corresponding tasks. The initial assessment was performed to get an understanding of the data quality in general and to identify methodology and tools for subsequent tasks.

## Data Structure

The entire dataset consists of four character-delimited files described below:

1. **Dish.csv**

This file contains dish names listed on the menu along with their respective pricing and chronology information. Each record represents a specific dish offered by a business and listed on the menu. Each dish has an identifier which uniquely identifies it and is referenced as a foreign key on other entities.

1. **MenuItem.csv**

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

1. **MenuPage.csv**

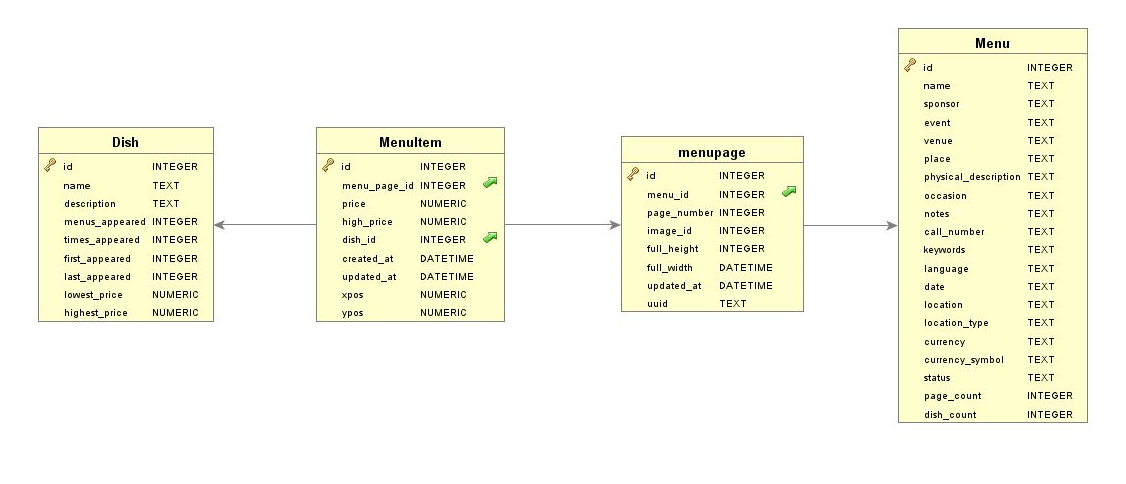
This file contrains menu page records. Each item is identified by a unique identifies and links a menu item with a menu. Additional information such as page photo image number and page dimensions also appear here. Ever record keep references of menu item identifier and menu identifier to link these entities together.

1. **Menu.csv**

This file contains all individual menus, each associated with a unique id. Each menu has an identifier which uniquely identifies it and is referenced as a foreign key on other entities. Associated data includes the occasion, venue and event information and chronological information such as created and updated dates and times. Other important fileds present in this file include the location where the menu is offered, the associated currency in use for the menu items, the language for the menu and the status, among others.

## Data Structure

The raw data was imported in an SQL-Lite instance and visualized using DB-Visualizer tool. The ER diagram generated from DB-Visualizer is shown the figure below:



The diagram shows entities and links for data objects present in the data set. Most of the raw data was imported as strings of characters. However, initial assessment showed following data types for the fields:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **Dish** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| Id | Integer |  | (10)9 | PK | N |
| Name | String | 1387 | X(1387) unicode |  | N |
| Description | String | 0 | X(1) |  | Y |
| Menus\_appeared | Integer |  | -(10)9 |  | N |
| Times\_Appeared | Integer |  | -(10)9 |  | N |
| First\_Appeared | Integer |  | (4)9 |  | N |
| Last\_Appeared | Integer |  | (4)9 |  | N |
| Lowest\_Price | Numeric | 2 | --------.99 |  | Y |
| Highest)Price | Numeric | 2 | --------.99 |  | Y |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **MenuItem** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| Id | Integer |  | (10)9 | PK | N |
| Menu\_Page\_Id | Integer |  | (10)9 | FK | N |
| Price | Numeric | 2 | ----.99 |  | Y |
| High\_price | Numeric | 2 | ----.99 |  | Y |
| Dish\_id | Integer |  | (10)9 | FK | Y |
| Created\_at | Timestamp(0)  With zone |  | YYYY-MM-DDBHH:MM:SS(0) Z |  | N |
| Updated\_at | Timestamp(0)  With zone |  | YYYY-MM-DDBHH:MM:SS(0) Z |  | N |
| Xpos | Numeric | 6 | -.999999 |  | N |
| Ypos | Numeric | 6 | -.999999 |  | N |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **MenuPage** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| Id | Integer |  | (10)9 | PK | N |
| Menu\_Id | Integer |  | (10)9 | FK | N |
| Page\_Number | Integer |  | ----.99 |  | Y |
| Image\_Id | String | 15 | X(15) |  | N |
| Full\_height | Integer |  | (4)9 |  | Y |
| Full\_width | Integer |  | (4)9 |  | Y |
| Updated\_at | STRING | 36 | X(36) [UUID] |  | Y |
| Uuid | Numeric | 2 | -.999999 |  | Y |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **Menu** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| id | INTEGER |  | (10)9 | PK | N |
| name | STRING |  | (10)9 | FK | N |
| sponsor | STRING |  | ----.99 |  | Y |
| event | STRING | 15 | X(15) |  | Y |
| venue | STRING |  | (4)9 |  | Y |
| place | STRING |  | (4)9 |  | Y |
| physical\_description | STRING | 36 | X(36) [UUID] |  | Y |
| occasion | STRING | 2 | -.999999 |  | Y |
| notes | STRING |  |  |  | Y |
| call\_number | STRING |  |  |  | Y |
| keywords | STRING |  |  |  | Y |
| language | STRING |  |  |  | Y |
| date | DATE |  |  |  | Y |
| location | STRING |  |  |  | Y |
| location\_type | STRING |  |  |  | Y |
| currency | STRING |  |  |  | Y |
| currency\_symbol | STRING |  |  |  | Y |
| status | STRING |  |  |  | Y |
| page\_count | INTEGER |  |  |  | Y |
| dish\_count | INTEGER |  |  |  | Y |

## Data Quality initial assessment

For an initial assessment, following data quality check were performed and respective violations were listed.

**Data.csv file:**

|  |  |  |
| --- | --- | --- |
| Field | Issue type | Description |
| Name | Case Standardization | The name do not appear in standard case. Some names are upper-case, some lower-case and other mixed-case.   * There are 426985 distinct dishes * There are 398443distinct dishes   e.g.  Dish.id = 1 , Dish.name = 'Consomme printaniere royal'  Dish.id = 397198, Dish.name = 'Consomme Printaniere Royal' |
| Name | Extra Spaces | There are extra space   * Leading dish names * Training disd names * In between dish names   e.g.  Dish.id = 131274, Dish.name = 'Consomme printaniere royal'  Dish.id = 397198, Dish.name = ' " " kidneys' |
| Name | Extra quotes | There are extra quotes in dish names  e.g.  Dish.id = 1788, Dish.name = 'Veuve Clicquot "Yellow Label"' |
| Name | Invalid characters | There are invalid characters like (!,@,#,{ etc.) in dish names  e.g.  Dish.id = 2839, Dish.name = 'E. & J. B. \*\*\*' |
| menus\_appeared | Plausibility | There are 2412 Dishes with menus\_appeared = 0  The field may be re-calculated from referenes |
| menus\_appeared | Correctness | There are differences in menus appeared and actual menu count for dish  e.g.  id =19, menu\_appeared = 16, actualy appeared = 15 |
| times\_appeared | Plausibility | Several 0 or negative values  -- 1 Dishes appeared -10 times ??? [MIN]  -- 11900 dishes appeared 0 times !!!  -- 372 dishes apperaed 19 Menus [MAX] |
| times\_appeared | Correctness | There are differences in times appeared and actual count for dish in menus  e.g.  id =17, times\_appeared = 535, actualy appeared = 536 |
| First\_appreaed | Correctness | Many dishes have first appesrance earlier than menu date |
| First\_appreaed | Plausibility | Several dishes have first\_appeared year later than last\_appreared year |
|  |  |  |

## Issues

Issues in Data:

**Dish:**

Cleansing in names required.

1> Same name in Upper and Lower Cases [OK] [Clustering pending]

e.g.

18958 Lobster salad

463235 Lobster Salad

2> Names appear winthin and without quotes [Multi space cleansing, quotes removal]

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. [OK]

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 15 menus instead of 16

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 validated (or may be overwritten) with minimum created\_at value in Menu\_Items assiciated with dish

Cleansing in last\_appeared.

same 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 occasion

standardize call number

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

## Use cases

## Data structure

## Fitness for use unrefined

## Fitness for use after cleaning

# Biblio

# Bibliography

|  |  |
| --- | --- |
| [1] | NYPL Labs, "What's on the menu Dataset," 16 June 2020. [Online]. Available: http://menus.nypl.org/data. |
| [2] | NYPL Labs, "Whats-On-The-Menu," June 2020. [Online]. Available: http://nypl.github.io/menus-api/. |
| [3] | David Huynh, Stefano Mazzocchi, Metaweb Technologies, Inc, "OpenRefine," October 2012. [Online]. Available: https://openrefine.org/. |
| [4] | SQLite Consortium, "SQLite," [Online]. Available: https://www.sqlite.org/index.html. |
| [5] | DBVis Software, "DbVisualizer: A universal database tool," DbVis Software AB, [Online]. Available: https://www.dbvis.com/. |