Final Project

End-To-End Data Cleaning Workflow – NYPL Data

by

Sudhir Behani

Bollam Raja Shekar

Table of Contents

[**Introduction and Overview** 3](#_Toc15405887)

[1. Dish.csv 3](#_Toc15405888)

[2. Menu.csv 3](#_Toc15405889)

[3. MenuItem.csv 3](#_Toc15405890)

[4. MenuPage.csv 3](#_Toc15405891)

[5. Missing Data 3](#_Toc15405892)

[Other Information 3](#_Toc15405893)

[**Initial Assessment of the dataset** 5](#_Toc15405894)

[**Data Cleaning methods and Process** 6](#_Toc15405895)

[Open Refine 6](#_Toc15405896)

[**Data Cleaning Results** 9](#_Toc15405897)

[Relation Database Schema 9](#_Toc15405898)

[ER Diagrams 9](#_Toc15405899)

[Integrity Constraint Checks 10](#_Toc15405900)

[Workflow Model 10](#_Toc15405901)

[Or2yw overall workflow 10](#_Toc15405902)

[Or2yw open Refine csv 10](#_Toc15405903)

[Overview of changes 11](#_Toc15405904)

[**Conclusion and Future Work** 12](#_Toc15405905)

# **Introduction and Overview**

## Dish.csv

Information about all the dishes from all the menus transcribed by the project are stored in data.csv. In this data file, a dish is represented by a row of values. Columns identify attributes of a dish. One of these attributes is an identifier, which identifies the dish. However, the identity of a dish appears to be based on the exact form of the string labeled "name." Thus, dishes with variant orthographic forms of their names, e.g. “half chicken”, “Half Chicken” and “chicken [half]”) are treated as separate entries with different identifiers.

## Menu.csv

Information about the menus as physical objects, including historical information about their origins, uses, and formats are stored in menu.csv

## MenuItem.csv

This is the largest data file. A "MenuItem" represents a single instance of a dish appearing somewhere on a menu page image. “MenuItem.csv” is useful as a mapping between multiple other data files/tables.

## MenuPage.csv

Information about individual pages of the menus is stored in "Menu.csv". Pages are modeled here as digital images produced as a result of digitization by the NYPL. Menus often have multiple pages.

## Missing Data

We note the presence of missing data points in the column “missing values.” However, there are also strings in the present data that point to missing information (e.g. “unknown” or “?”), which do not formally appear as null values.

## Other Information

The data in these files comes from several different sources. We reflect this in the “generated by” category of the tables below. Some of the data is supplied by “volunteer transcribers,” by which we mean people who have participated in the project through the What’s On the Menu? site. Some of the information is generated by “web application.” This means that some of this information was automatically created as the database supporting the application was constructed and populated (e.g., various ids); some is created as the web application runs (e.g. timestamps as data values are updated). Finally, a lot of information is generated from “NYPL metadata.” This metadata comes from many places and reflects the long history of the project and the many parts of New York Public Library involved in it. Much of the data “supplied by NYPL metadata” in the menu spreadsheet is from the catalog cards made by Frank E. Buttolph in the early twentieth century.

# **Initial Assessment of the dataset**

# **Data Cleaning methods and Process**

## Open Refine

1. Menu

Activities performed

* Trim leading and trailing white spaces:
  + Columns: sponsor, event, place, physical\_description, notes, location, Call\_number
* Collapse consecutive white spaces:
  + Columns: sponsor, event, place, physical\_description, notes, location,
* Convert columns to number type:
  + Columns: page\_count
* Convert columns to date type:
  + Columns: date
* Create a new column date\_timeless so that date in YYYY-MM-DD format without time is recorded. A new column is created so that no values are lost by updating date column.
* Remove special characters including semicolon at the end of the string

Note: Can’t remove ; (semicolon) character in the middle of string as this

signifies multiple options- example multiple sponsors, event, venue, place etc.

Columns: All the columns with string data type

* Convert all values in columns with string data type to upper case so that no confusion in case creates issues in finding similar values.
* After removing special characters again perform white spaces removal step for all the string data type columns
* Clusters columns based on methods - key collision and keying function is fingerprint.

Note: This keying function fails to cluster all the similar values to the same cluster. Visually identifying those clusters and then names of similar looking clusters were manually added in the New Cell value.

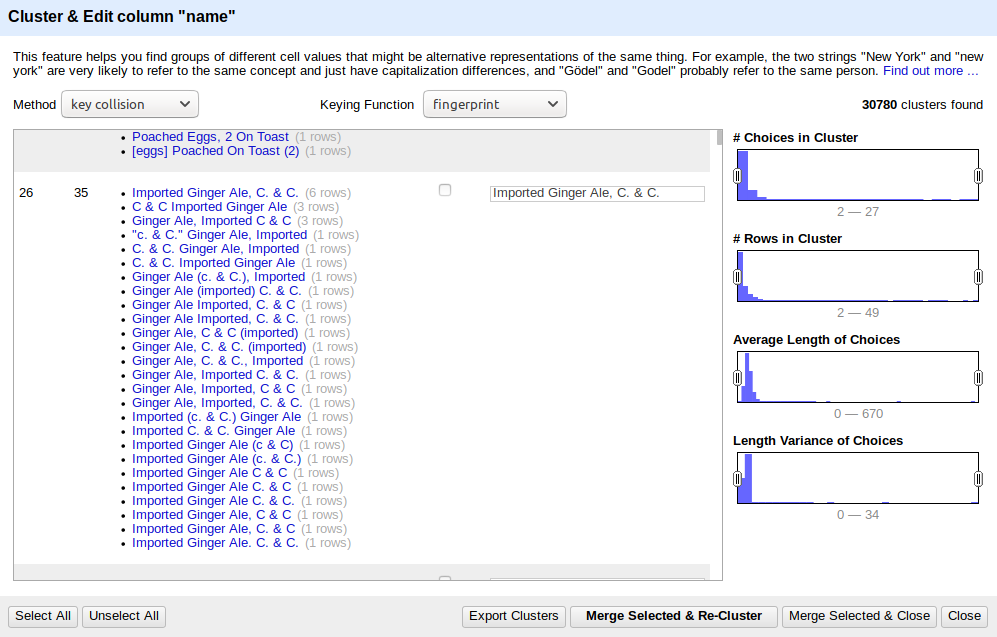
Difficulties:

* There occasions when “,” (comma) is interchangeably used for multiple items and additional information after comma. This needs to be taken care separately.

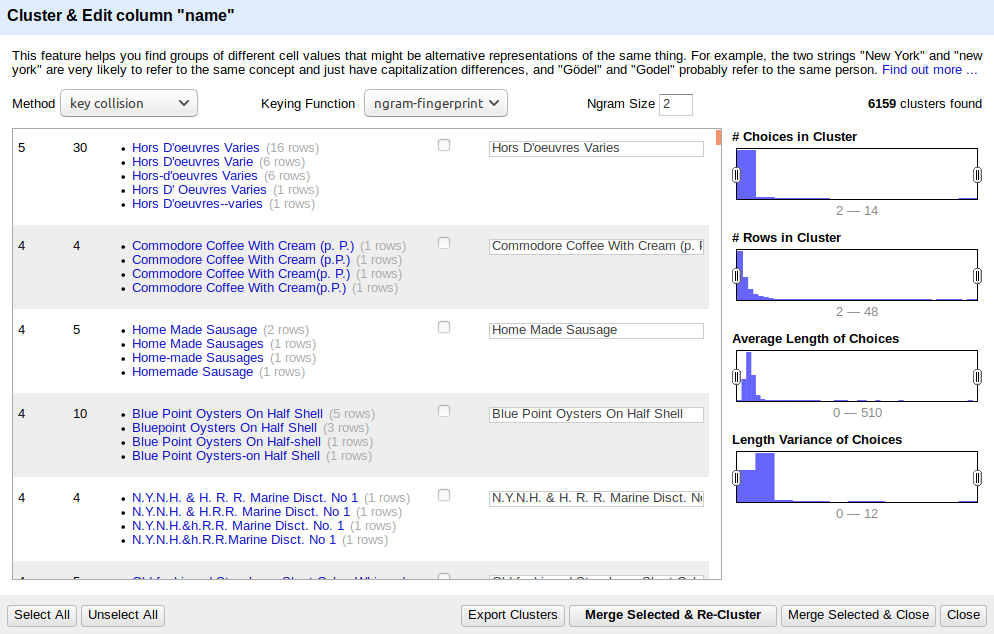
1. Dish

Activities performed:

* Trim and Collapse white spaces in “name” column.
* Convert column “name” to title case.
* Make a facet and perform the cluster operation using the “key-collision” method and “fingerprint” as shown below.



* Repeat step c with “ngram-fingerprint” as shown below.



* Remove “description” column, since most of the entries are empty.

1. MenuPage
2. MenuItem

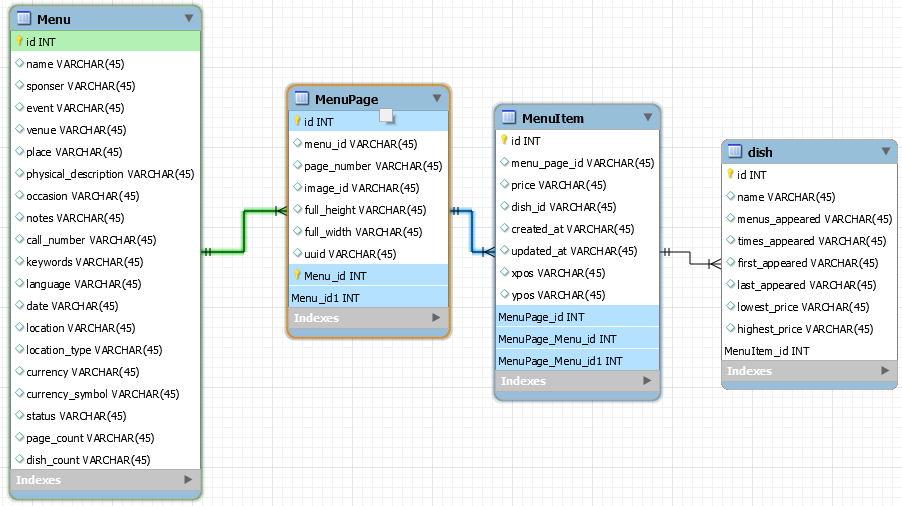
Following data cleaning steps were performed on the Dish dataset.

* “high\_price” column is removed
* “created\_at” is transposed into “toDate()”
* “updated\_at” is transposed into “toDate()”

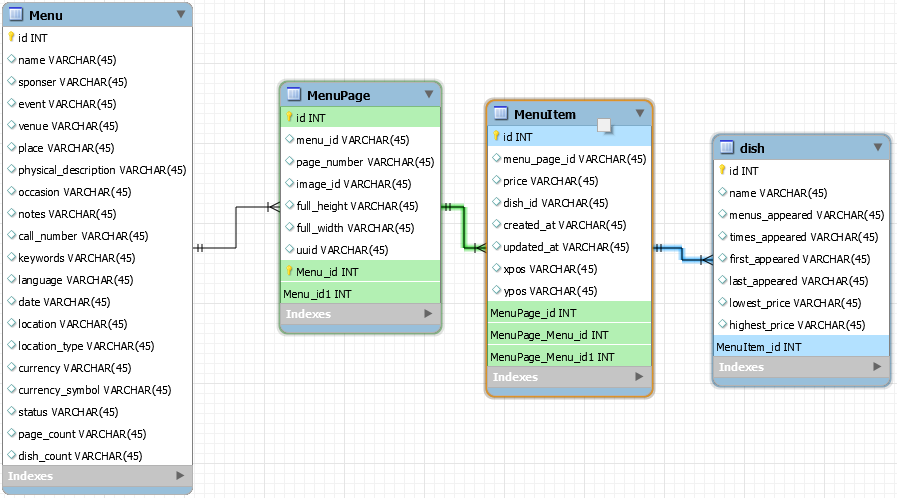
# **Data Cleaning Results**

## Relation Database Schema

### ER Diagrams



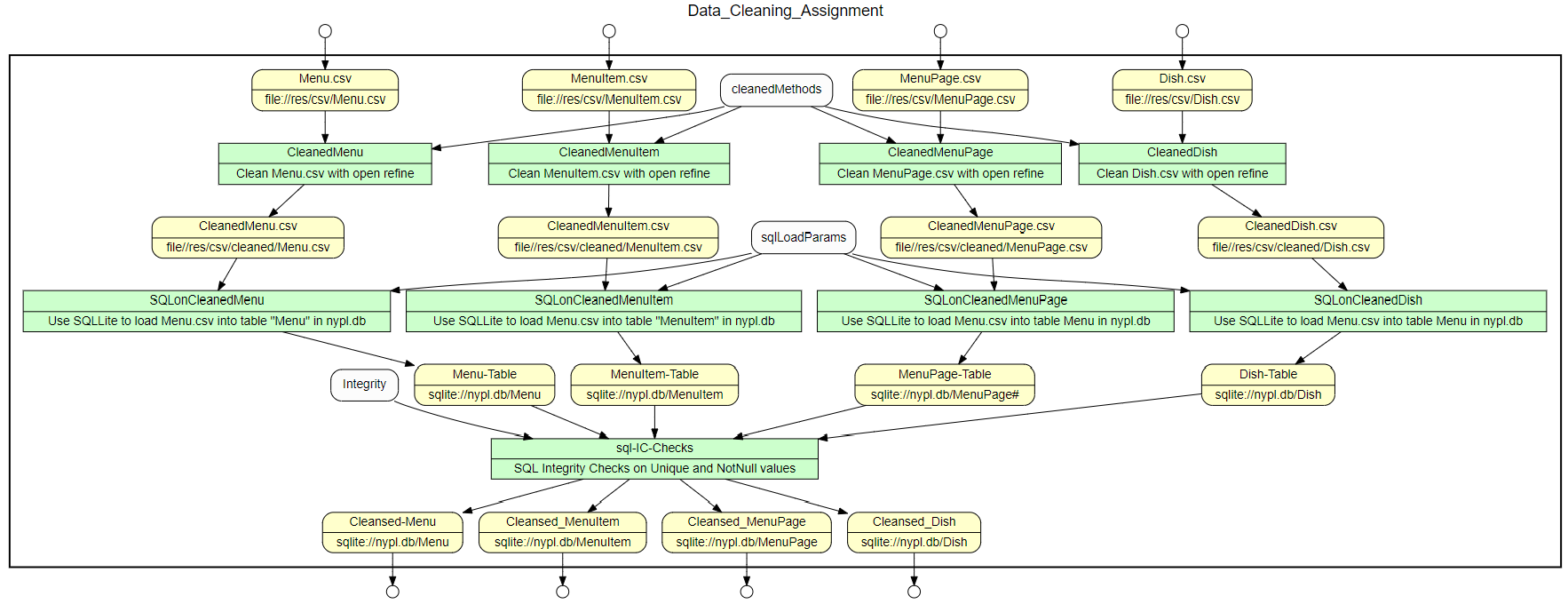
<<<Add text >>>



### Integrity Constraint Checks

## Workflow Model

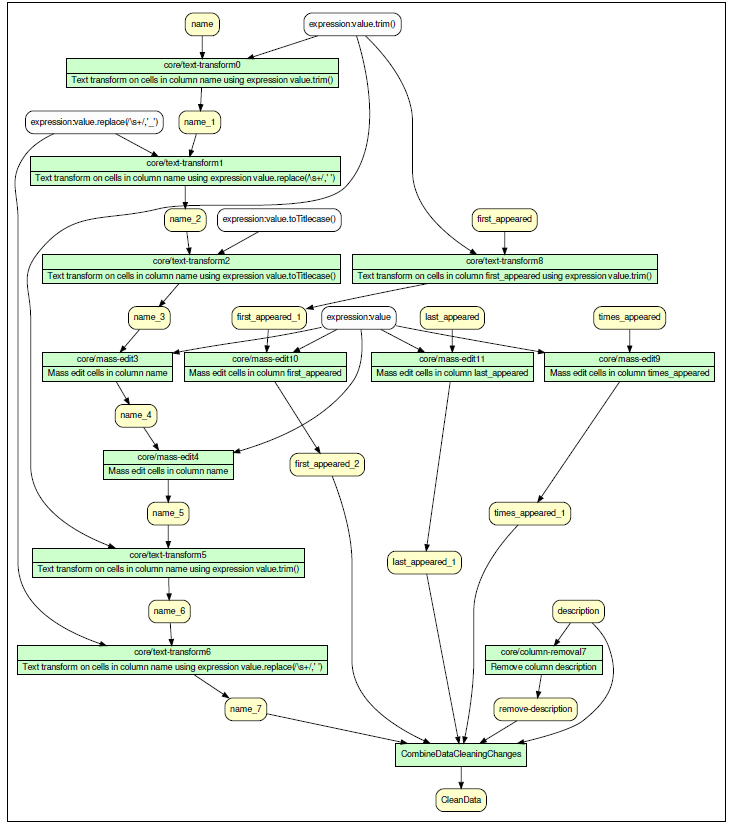
### Or2yw overall workflow



### Or2yw open Refine csv

Following graphs are generated using or2yw tool based on the recipes generated from the openrefine too.

1. Dish.csv



1. Menu.csv

## Overview of changes

# **Conclusion and Future Work**