CS 513 – FINAL PROJECT – PHASE I

TEAM 169 - CHIRANJEEVI (CHIRU) KOILOTH, CHUN CHIEH (WILLY) CHANG, AARON BARRIE

1. DATASET CHOSEN

Our team has chosen the New York Public Library Historical Menu dataset[1], also known as the New York Public Library "What's on the menu?" dataset.

2. DESCRIPTION OF DATASET

The Historical Menu dataset is delivered as a set of four CSV files where each file is representative of an entity in the dataset:

1. Menu.csv
2. MenuPage.csv
3. Dish.csv
4. MenuItem.csv

Overall, this dataset describes a collection of menus and dishes that exist within a crowdsourced data set produced by the New York Public Library. Each menu is made of a collection of Menu Pages. Each Menu Page is made of a collection of Menu Items. Each Menu Item relates a Menu Page to a Dish, along with pricing information.

The Menu, Menu Page, and Menu Item entities could be considered hierarchical. Each menu represents a single physical document. Each Page belongs to precisely one Menu. Each Menu Item belongs to precisely one Menu Page. The Dish entity is the only entity that can belong to one or many Menu Items, and therefore indirectly belong to one or many Menu Pages, and finally could indirectly belong to one or many Menus.

This dataset shows records that date back as early as 1851 and as recent as 2013. There are years included in the menu dates that precede 1851 but are themselves likely to be data errors.

MENU.CSV

In terms of how the dataset is distributed, Menu.csv can be considered as the root entity within this dataset. Each record in this dataset represents a single menu document. Each document/row in this dataset corresponds to a single physical menu document with the following information associated with it:

- Unique ID for the menu.
- Name of the restaurant/organization serving the menu.
- The type of menu, e.g., Breakfast, Lunch, Dinner, or some specific event such as a menu for a specific single-night event.
- The type of venue, e.g., Commercial venue, Social Hall, etc.
- Location of the venue, with varying degrees of specificity.
- Description of the menu, e.g., document format, dimensions
- Special Occasion, when appropriate

[1] https://uofi.app.box.com/s/whvfh9jio38ck0m9gz58s31srx8iwg4i/folder/159094620210

- Miscellaneous notes, generally describing the material or condition of the menu.
- A document Call Number
- Date of menu publication
- Location Name
- Currency Name
- Currency Symbol
- Menu Digitization Status (complete, or under review)
- Page Count
- Dish Count

While we may only have one record in this file for each document, there may be multiple documents for the same restaurant/organization, e.g., one restaurant could have multiple documents for Breakfast, Lunch, and Dinner service, and additionally could have different menus for the same service published on different dates.

Note: some fields have been excluded from the list due to being entirely empty, or at this point seeming to be a redundant copy of another field. These fields are represented in the data model diagram in this section, and further details will be described in the Phase II Project submission after additional analysis and cleanup has been performed.

MENUPAGE.CSV

The MenuPage.csv file describes the pages that comprise a Menu document. Each record represents a single face of a single page in a menu. Each menu should have at least one page but can be made up of many pages. Each page has the following information associated with it:

- Unique Page ID.
- Reference to the Unique ID for the menu to which this page belongs.
- The sequence number for this page.
- A unique Image ID referencing an Image file external to this dataset.
- The image height (in pixels).
- The image width (in pixels).
- A Universally Unique ID (UUID) reference for the Image file external to this dataset.

Each menu page should have one row in this dataset, and no page should have more than one row in this dataset.

DISH.CSV

The Dish.csv file describes the types of dishes that are available as Menu Items across the Menus in this dataset. This is intended to be an entity that can be referenced by multiple menus, with each row representing a conceptually distinct Dish being served within the overall dataset. Each dish has the following information associated with it:

- Unique Dish ID.
- Dish Name.
- Count of menu appearances for this dish in the dataset (unique instances across all Menu entities).
- Total number of appearances of this dish in the dataset (unique instances across all Menu Item entities).
- Year of first appearance in this dataset
- Year of last appearance in this dataset

- Lowest price of this dish across all Menu Item entities in this dataset.
- Highest price of this dish across all Menu Item entities in this dataset.

Like a Manu, the Dish entity does not directly relate itself to another entity in the dataset. While a Dish can be directly related to a Menu Item, and therefore be indirectly related to a Menu Page and Menu, there is no attribution present on the Dish entity itself that describes such a relationship.

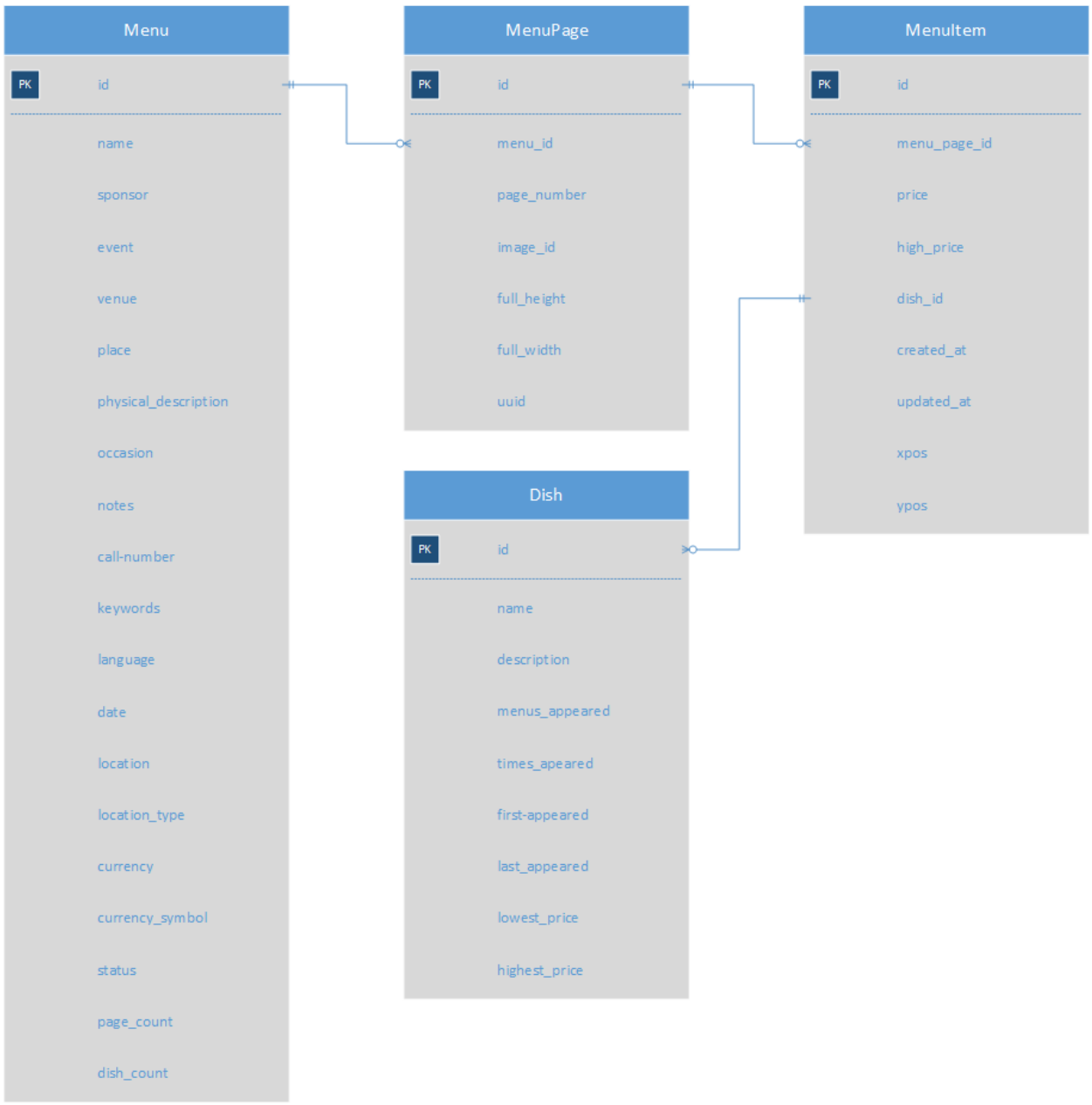
MENUITEM.CSV

The MenuItem.csv file describes a menu item (i.e., Dish) that is presented on a menu page. It effectively relates a Dish entity to a Menu Page entity, associating it with a Price and a specific position on the page. The following information is associated with a Menu Item:

- Unique Menu Item ID
- Reference to the Unique ID for the Menu Page to which this item belongs.
- The Price of this dish (in the currency units associated with the Menu to which this item belongs).
- The Higher Price of this dish (in the currency units associated with the Menu to which this item belongs), presumably if a dish can be ordered in different portion sizes.
- Record Creation Timestamp.
- Record Last Update Timestamp.
- Horizontal Position on the page (on a 0 to 1 scale).
- Vertical Position on the page (on a 0 to 1 scale).

DIAGRAM

The following diagram shows the entities, attributes, and identified relationships within the selected data set, acquired from the course Dropbox folder provided in the project description.



3. USE CASES

U0 – ZERO CLEANING

The team wishes to build a web-based application to deliver the contents of this dataset for an embedded system where the image references indicated in the MenuPage entity of this dataset are unavailable for distribution. Each menu can be identified from a navigable list that provides search and filtering based on:

- Menu Name
- Location
- Special Occasion
- Date of Publication

- Keyword Search for Notes

Clicking on a menu will open a navigable document that has been reconstructed based on the information from the Menu Page, Menu Item, and Dish entities.

The Menu Page entity will provide the Menu Page ID and page dimensions, which will be scaled to fit the current browser window.

All Menu Items associated with a given Menu Page will be plotted using a textbox with an origin point located at the relative X/Y described by xpos/ypos. This textbox will include:

- Dish Name (from the Dish entity)
- Dish Price (from the MenuItem entity)
- Dish Alternate Price (from the MenuItem entity, when available)
- Dish Prices will be decorated with a Currency Symbol from the Menu entity

Additionally, this application will support the ability to filter the list of Menu's based on a search for Dish names. If a menu is selected when this filter is used, then the application will automatically navigate the user to the page on which that Dish is found, with a highlight placed around the location of the Menu Item textbox on that page.

Any data points presented in the original data set will be done as-is. Any data points that are inaccessible due to integrity constraint violations will result in a "Missing Page" notice to the end user in the even to of a missing page reference, or "Missing Data" notice to the end user in the event of a missing dish reference. When currency markers are unavailable, the prices will be listed as-is without currency designation.

U1 – MAIN USE CASE

The team wishes to extend the functionality of the application described in U0 by adding the ability to offer additional Organization- and Dish-based analysis.

To offer Organization-based analysis, the data points for Name/Sponsor on the Menu entity need to be appropriately clustered to accurately present a comprehensive view of how the menu characteristics for a single organization change over time.

To offer Dish-based analysis, the data points for the Dish Names from Dish.csv and prices from MenuItem.csv need to be cleaned to provide an accurate view for how the popularity of specific dishes changed over time, and how price trends for dishes change over time.

U2 – NEVER ENOUGH USE CASE

The team has been contracted by Microverse to build an automated content generation system for their brand new fully immersive virtual reality dining experience. The client has requested that adequate metadata be provided to their Stable Diffusion-based Virtual Waiter experience, to generate a VR-based experience where a user can select a menu and then be brought into an immersive simulation of the restaurant environment, where a location- and period-appropriate waiter would patiently describe the contents of the menu, including any dish of the user's interest.

To do this, not only would the application rely on the cleanup described in the U1 Main Use Case, but would additionally require accurate information about:

- Physical Location of the dinner service
    - While there are sometimes addresses present in the location data for the Menu entity, most of the data points are too generic for a precision location determination. Additionally, some records describe menus that are served on transatlantic journeys (see: Hamburg Amerika Linie).
- Socioeconomic class standards for the dining experience
    - Going back to the same Hamburg Amerika Linie example above, while the pricing data points could be cross-referenced against comparable pricing for other menus of the same year., there would be a lack of confidence to determine if this menu was being presented to first-class passengers or passengers in a lower/steerage class.
- Dish descriptions
    - A key data point in the accurate delivery of the authentic waiter experience is a description of the dish. While the Dish entity contains a description attribute, not a single record contains any information populated within.

4. DATA QUALITY PROBLEMS

OBVIOUS DATA QUALITY PROBLEMS

After a preliminary review of the source data set, there are a variety of data quality problems that are clearly identifiable. While a deeper review will be completed ahead of the Phase II project delivery, some obvious cases will be described below.

MISSING DATA VALUES

There are a variety of attributes across entities where some or all records have a null or empty value. These missing data points will be a hurdle for data processing and analytic tasks. Such empty attributes include:

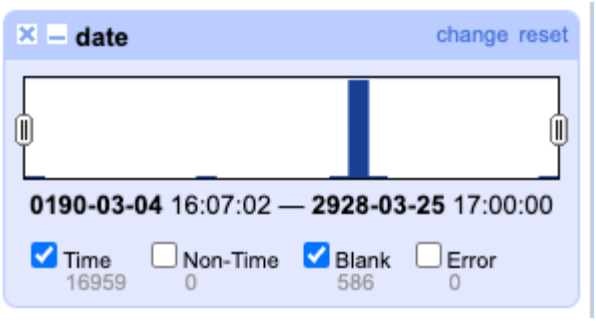
- Menu.csv
    - Name
    - Sponsor
    - Event
    - Venue
    - Phsyical Description
    - Occastion
    - Notes
    - Call Number
    - Keywords
    - Language
    - Location
    - Location Type
    - Currency
    - Currency Symbol
- MenuItem.csv
    - Price
    - High Price
- MenuPage.csv
    - Full Height

    - Full Width
- Dish.csv
    - Description

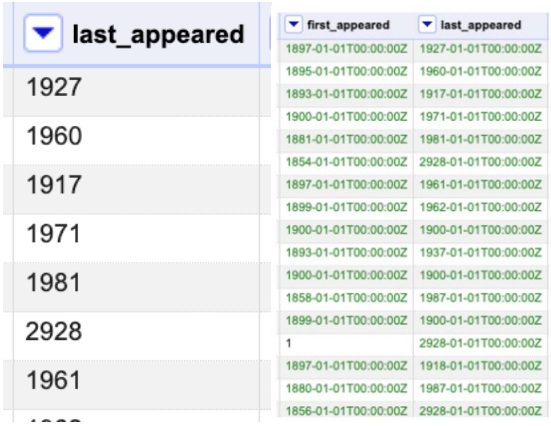
DATE TIME RANGE CONCERNS

The Date attribute within the Menu entity includes some values that are clear outliers from the expected date ranges. Similar unexpected outliers occur within the last_appeared attribute of the Dish entity.

This screenshot shows the outliers for the Menu.date attribute.



The following screenshot shows outlier data for the Dish.last_appeared attribute.



last_appeared
1927
1960
1917
1971
1981
2928
1961

first_appeared	last_appeared
1897-01-01T00:00:00Z	1927-01-01T00:00:00Z
1895-01-01T00:00:00Z	1960-01-01T00:00:00Z
1893-01-01T00:00:00Z	1917-01-01T00:00:00Z
1900-01-01T00:00:00Z	1971-01-01T00:00:00Z
1881-01-01T00:00:00Z	1981-01-01T00:00:00Z
1854-01-01T00:00:00Z	2928-01-01T00:00:00Z
1897-01-01T00:00:00Z	1961-01-01T00:00:00Z
1899-01-01T00:00:00Z	1962-01-01T00:00:00Z
1900-01-01T00:00:00Z	1900-01-01T00:00:00Z
1893-01-01T00:00:00Z	1937-01-01T00:00:00Z
1900-01-01T00:00:00Z	1900-01-01T00:00:00Z
1858-01-01T00:00:00Z	1987-01-01T00:00:00Z
1899-01-01T00:00:00Z	1900-01-01T00:00:00Z
1	2928-01-01T00:00:00Z
1897-01-01T00:00:00Z	1918-01-01T00:00:00Z
1880-01-01T00:00:00Z	1987-01-01T00:00:00Z
1856-01-01T00:00:00Z	2928-01-01T00:00:00Z

Based on our preliminary review we expect the valid dates to exist between 1851 and 2013.

INCONSISTENT DATA FORMATS

There are instances where representations of the same data types are not consistently formatted. Menu.date has a variety of inconsistent date formats (including yyyy/mm/dd and mm/dd/yyyy). Additionally, the UUID values for the MenuPage entity have inconsistencies in casing. While we anticipate that these values are unique per page, and therefore likely to be non-repeating, there are inconsistencies with how the string representation of the UUID appears in the MenuPage.csv file.

510d47db-491e-a3d9-e040-e00a18064a99
510D47DB-491F-A3D9-E040-E00A18064A99
510d47db-4920-a3d9-e040-e00a18064a99

SPECIAL CHARACTERS

Inconsistent use of special characters: The use of brackets [] in some entries in the event column isn't consistent, with some entries encapsulating the event in brackets and others not. This inconsistency needs to be addressed.

INCONSISTENT NAMING

Within both the Dish names in Dish.csv and the Location names in Menu.csv (as well as other string values within the overall dataset) we have multiple representations of the same value that likely refer to the same entity. For example, "Norddeutscher Lloyd Bremen" and "Norddeutscher Lloyd BREMEN" are likely the same but are written differently.

- Norddeutscher Lloyd Bremen (687 rows)
- Norddeutscher Lloyd Bremen (45 rows)
- Norddeutscher Lloyd Bremen (41 rows)
- Norddeutscher Lloyd Bremen (7 rows)
- Norddeutscher Lloyd, Bremen (2 rows)
- Bremen Norddeutscher Lloyd
- Norddeutscher Lloyd Bremen;

INAPPROPRIATE ZERO VALUES

Within MenuItem.csv, we see prices that are listed as 0 that likely should be a non-zero value. Additionally, the prices represented in the lowest_price and highest_price attributes of Dish.csv show zero values.

lowest_price
0.2
0.1
0.25
0.25
0.0
0.0

WHY CLEANING IS NECESSARY FOR U1

The preceding examples of data errors will have a clear negative impact on the delivery of the application described in U1 – Main Use Case.

If we intend to provide some measure of accuracy for analysis of pricing and menu offerings within a single restaurant/location/organization, we will need to unify/cluster the representations of the corresponding attribute names that the application would rely on. Importantly, we will need to clean up the clusters of similar representations of names of Menu.location and Dish.name to accomplish our stated goal of providing a comprehensive view of price trends/popularity trends/etc. across common Dish and common Location references.

If we want to accurately present Price trend history for specific dish items, we will additionally need to correct for invalid price data where we should be excluding zero-value Dish prices from our consideration.

There are calculated range data points on the Dish entity that may need to be recalculated either at runtime, or as part of the source data set (by way of data cleanup) to correct for low/high price ranges for a dish across the source data set once the zero-value dish instances are removed from consideration.

5. INITIAL PLAN FOR PHASE II

S1 – REVIEW

Aaron will own the summary of the data quality problems that we were able to see via OpenRefine. We will also work backwards from U1 and check to see what are the vital data quality promises we need to keep in order to ensure successful delivery of U1. This will result in a list of IC-Violation checks as well.

Assignment: Aaron

Timeline: Executed July 10th through July 12th

S2 – PROFILE

To profile the data set we will use Open Refine first and check to see if we can dive deep into the data and understand what the deeper data quality issues are.

Then we will load the dataset into a SQL dataset and try to make sure that the IC-Violation checks are not broken.

Assignment: Chiru

Timeline: July 13th through July 16th

S3 – PERFORM

We will use data cleaning tools such as OpenRefine, Python, Pandas, and NumPy to clean the data. Willy will handle missing values, standardize and correct inconsistent data entries, and identify and deal with outliers.

Assignment: Willy, Aaron, Chiru

Timeline: July 17th through July 24th

S4 – DQ CHECKING

We are going to compare the new data set with the old data set using OpenRefine first. The problems that we had caught in section 1 should not be there in the new data set.

We will compile a quantitative measure of the number of columns changed, and we will run this through our IC-Violation list to make sure the data is as clean as is required for U1.

Assignment: Willy, Aaron, Chiru

Timeline: July 25th through July 28th

S5 – DOCUMENT AND QUANTIFY CHANGE

To document all of these changes we will compile all of the measured differences that we have made in the new data set into a report and explain how it pertains to the use case.

We will also talk about a YesWorkflow that will ensure we can do this repeatedly over time.

Assignment: Chiru, Aaron

Timeline: July 29th through July 30th