**CS513: Theory & Practice of Data Cleaning Project**

End-to-End Data Cleaning Workflow

MCS-DS program at the University of Illinois at Urbana-Champaign – Fall 2018

Abstract: The report describes a data cleaning workflow using the New York Public Library Rare Books Division 45,000 historical menus dataset has an example file to demonstrate various techniques. Over the course of the project we will use OpenRefine, SQLite, and YesWorkFlow to clean and organize the dataset.

**Team**

Brad Ballard - bjb3@illinois.edu

Dhanendra Singh - disingh2@illinois.edu

Jacob Rettig - jrettig2@illinois.edu

**1. Dataset Overview and Initial Assessment**

We use tools introduced in CS513 to clean and prepare a sample dirty dataset (NY menus). Along the way we will document the steps from dirty to clean. The sections that follow describe the specific steps in detail: (2.) Data cleaning with OpenRefine, (3.) Develop Rational Database Schema, and (4.) Create a Workflow Model.

An initial assessment of the New York Public Library dataset (referred to now forward as NY menus dataset) is over 45,000 historical menus. The majority of these were organized by Frank E. Buttolph (reference1) around 1900-1921. The dates on the menus range from the 1850s to 2010s. The data contains information on restaurant menu, as well as,

**2. Data cleaning with OpenRefine**

here

**3. Develop Rational Database Schema**

here

**4. Create a Workflow Model**

here