Paul Nguyen, paultn2@illinois.edu

Shrirang Bagdi, sbagdi2@illinois.edu

Fengdi Liu, fengdil2@illinois.edu

Project Plan – Chicago Food Inspections Data Curation

1. Overview

The purpose of this project is to design and implement an end-to-end data curation workflow using the Chicago Food Inspections dataset supplemented with related datasets such as Business License and 311 Sanitation Code Complaints from the City of Chicago Open Data Portal. Food safety is a critical area of public health, and restaurant inspections play a key role in maintaining safety standards. However, inspection data often contains inconsistencies, unstructured fields, and missing values that limit its utility for analysis.

The objective of our project is to curate an integrated, high-quality dataset that can be used to explore patterns of restaurant inspection failures. By aligning our workflow with the data lifecycle framework, we will demonstrate the processes of acquisition, profiling, cleaning, integration, metadata creation, workflow automation, provenance capture, and dissemination. While the central deliverable is a curated dataset and reproducible workflow, we will also conduct exploratory analyses to validate our work. For example, examining the most frequent violations and evaluating whether license status or sanitation complaints correlate with failure rates.

This project aligns with course objectives because it emphasizes the principles of data curation, reproducibility, transparency, and documentation. It will allow us to practice skills introduced in class while producing outputs that are reproducible and reusable by others.

2. Plan (Stages Aligned with Data Lifecycle)

1) **Data Acquisition and Ethical Review**

   The first stage of our workflow will involve acquiring datasets from the City of Chicago Open Data Portal. The primary dataset is the Food Inspections dataset, which records inspection results, violation details, and outcomes. To provide additional context, we will also acquire the Business Licenses dataset and the 311 Service Requests – Sanitation Code Complaints dataset. Because these datasets are publicly available under open-data licensing, we do not anticipate significant legal or ethical barriers; however, we will document licensing terms carefully to ensure compliance. This stage will conclude with a record of data sources, access methods, and ethical considerations.

2) **Data Profiling and Quality Assessment**

   Once the datasets are collected, we will conduct profiling to assess their structure and quality. This includes examining the extent of missing values, detecting duplicate records, and identifying inconsistencies across fields such as violation descriptions, inspection types, and location identifiers. We will evaluate data quality using standard dimensions: completeness, accuracy, consistency, and timeliness. Deliverables for this stage will include a profiling report summarizing findings and outlining strategies for cleaning and integration.

3) **Data Cleaning & Integration**

   The next stage focuses on preparing the datasets for analysis. We will standardize inconsistent values, such as city names, date formats, and inspection type codes. Unstructured violation text will be normalized into structured categories (e.g., pest control, sanitation, food storage), improving usability. Integration will involve linking inspections to license records and sanitation complaints using unique identifiers such as license numbers, addresses, or establishment names. This stage will produce a clean and integrated dataset that serves as the foundation for subsequent analysis.

4) Data Modeling & Metadata

To formalize relationships between entities, we will develop an entity-relationship (ER) model that links inspections, establishments, and complaints. We will also generate metadata that conforms to recognized standards such as schema.org or DataCite, ensuring interoperability and discoverability. A comprehensive data dictionary will describe variables, permissible values, formats, and units. Deliverables from this stage will include an ER diagram, metadata files, and a data dictionary.

5) Workflow Automation & Provenance

Reproducibility will be achieved through automated workflows implemented in Python. Provenance will be tracked using GitHub commit history and annotations through tools such as YesWorkflow, providing transparency in data transformations. To ensure that others can replicate our environment, we will provide a requirements.txt file within our GitHub repository. This stage will produce a fully automated workflow with transparent provenance documentation.

6) Packaging & Dissemination

The final stage of the project will focus on packaging and dissemination. All scripts, curated datasets, metadata, and documentation will be archived in a GitHub repository. A README.md file will provide step-by-step instructions for reproducing the workflow and analysis. Finally, we will prepare a narrative report summarizing the workflow, lessons learned, and potential directions for future research. Deliverables will include a reproducible repository, complete documentation, and the final project report.

3. Data Sources

Our project will utilize three datasets from the City of Chicago Open Data Portal, each of which provides complementary information relevant to food safety and restaurant inspections:

- Food Inspections – Contains detailed records of inspections conducted by the Chicago Department of Public Health. This dataset includes establishment information, inspection dates, results (e.g., Pass, Fail, Pass with Conditions), and free-text violation descriptions. It provides the primary data for assessing compliance with food safety regulations.
- Business Licenses – Includes licensing information for all businesses operating in Chicago, including legal names, "doing business as" (DBA) names, addresses, license categories, application dates, expiration dates, and license status. This dataset will allow us to evaluate whether license type or status is associated with inspection outcomes.
- 311 Service Requests – Sanitation Code Complaints – Contains records of citizen-reported sanitation complaints submitted through Chicago's 311 system, including request types, locations, and resolution statuses. Since sanitation complaints may precede or correlate with inspection failures, this dataset provides valuable context for understanding risk factors.

All three datasets are publicly available, regularly updated, and distributed under open-data licensing through the City of Chicago Open Data Portal. Their availability and documentation make them suitable for an academic data curation project that emphasizes transparency, reproducibility, and ethical use.

4. Team Roles & Responsibilities

- **Shrirang Bagdi (Data & Acquisition):** Responsible for identifying and collecting datasets from the City of Chicago Open Data Portal, reviewing licensing terms, and documenting ethical and legal considerations. Member A will also perform dataset profiling, including assessing completeness, consistency, and potential issues, and will produce documentation of the data sources used in the project.

- **Paul Nguyen (Processing & Workflow):** Responsible for data cleaning, transformation, and integration across the Food Inspections, Business Licenses, and 311 Sanitation Complaints datasets. Member B will also develop reproducible workflow scripts in Python and implement provenance tracking through GitHub commits and workflow annotation tools (e.g., YesWorkflow or Snakemake). This ensures that data processing steps are transparent and reproducible.
- **Fengdi Liu (Metadata & Reporting):** Responsible for metadata creation, including development of an entity-relationship (ER) model and a comprehensive data dictionary. Member C will also coordinate the drafting of the proposal, progress report, and final report, ensuring APA-style citations and alignment with course concepts.

All members will jointly review deliverables, contribute to exploratory analyses, and co-author written reports. This collaborative structure ensures that the project benefits from multiple perspectives while maintaining clear ownership of specific tasks.

5. Timeline

Week	Task	Responsible	Deliverable
1–2	Define research question, select datasets, assign roles	All	Proposal (9/15)
3–4	Acquire datasets, perform profiling, quality assessment	Shrirang B	Dataset documentation
5–6	Clean inspection data (dates, city names, violation text)	Paul Nguyen	Cleaned dataset

7–8	Integrate with licenses/complaints; draft ER model	Shrirang B & Paul N.	Draft workflow + ER diagram
9	Metadata creation, data dictionary	Fengdi Liu	Metadata package
10	Provenance setup, GitHub repository	Shrirang B	Automated workflow
11	Status update (progress report)	Fengdi L. (All support)	Progress Report (10/27)
12–13	Workflow refinement, final curated dataset	Paul N.	Curated dataset
14	Documentation polish, packaging	Fengdi Liu	Archive-ready repo
15	Final report writing	All	Final Report (12/10)

6. Constraints

- **Data volume**: The combined Food Inspections, Business Licenses, and 311 Sanitation Complaints datasets can exceed 100 MB and include millions of records. Handling this scale requires efficient processing strategies, such as chunked reading, indexing, or using database-style queries. While the data size is manageable with modern tools, it will necessitate careful consideration of computational resources to ensure smooth workflow execution.
- **Data quality**: A significant challenge lies in the violation descriptions within the Food Inspections dataset, which are free-text and highly inconsistent. This unstructured format complicates categorization and analysis. To address this, we plan to normalize violation text into broader categories (e.g., sanitation, pest control, food storage), but this process will involve subjective decisions and may introduce classification errors.
- **Integration challenges**: Linking inspections to business licenses and sanitation complaints requires reliable identifiers. However, inconsistencies in business names, addresses, and license numbers may make exact matching difficult. We will need to develop a combination of

deterministic and fuzzy matching strategies to integrate records, and even then, some connections may remain ambiguous or incomplete.

- **Scope**: While predictive modeling of inspection failures could be valuable, the primary focus of this course project is on curation and workflow reproducibility. For this reason, any predictive analysis will be exploratory only and limited to illustrating the usefulness of the curated dataset, rather than developing a fully validated predictive model.

7. Gaps

- **Provenance tool choice**: Still deciding between YesWorkflow, Snakemake, or custom logging.
- **Metadata standard**: Will finalize between schema.org JSON-LD and DataCite XML.
- **Analysis scope**: Need clarification from course staff whether predictive analysis is encouraged or optional.

Sources

City of Chicago. (2024, July 26). *311 Service Requests – Sanitation Code Complaints – Historical* . City of Chicago Data Portal. https://data.cityofchicago.org/Service-Requests/311-Service-Requests-Sanitation-Code-Complaints-Hi/me59-5fac

City of Chicago. (2025, September 14). *Business licenses*. City of Chicago Data Portal. https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses/r5kz-chrr

City of Chicago. (2025, September 14). *Food inspections*. City of Chicago Data Portal. https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5

City of Chicago. (2025, September 14). *Business owners*. City of Chicago Data Portal. https://data.cityofchicago.org/Community-Economic-Development/Business-Owners/ezma-pppn

DataCite. (n.d.). *DataCite*. https://datacite.org/

Bagdi, S. Liu, F. Nguyen, P. (2025, September 14). *598-Chicago-Food-Inspections*. GitHub. https://github.com/paul-nguyen-1/598-Chicago-Food-Inspections

Schema.org. (n.d.). *Schema.org*. https://schema.org/

YesWorkflow. (n.d.). *YesWorkflow*. https://yesworkflow.com/