Critical Acclaim vs. Reader Popularity:

Curating an Integrated Dataset of Literary Awards and Book Reception

Author: Somnath Saha (*somnath4@illinois.edu*)

1. Overview

The publishing world often distinguishes between "*critically acclaimed*" literature and popular "*bestsellers*," but the relationship between these two categories is complex and not always well-defined. This project aims to address this gap by planning and implementing an end-to-end data curation workflow to produce a high-quality, integrated dataset for analyzing the correlation between major literary awards and public book popularity.

The primary goal is not to perform a deep statistical analysis, but to master the curation process itself: acquiring heterogeneous data, assessing its quality, cleaning and integrating it, and packaging it in a reproducible and reusable format. The final curated dataset will be structured to help answer research questions such as:

- Does winning a major literary award correlate with a significant, measurable increase in reader ratings or sales performance?
- How much overlap exists between books that win prestigious awards and those that appear on bestseller lists?
- Can a book's popularity metrics (e.g., number of ratings) before an award nomination predict its likelihood of winning?

This project will provide hands-on experience like API-based data acquisition, data modeling, quality assessment, data cleaning, workflow automation, and metadata creation, directly

applying the core concepts of the *CS598 - PSL* course.

2. Plan

The project will follow a data lifecycle model, encompassing stages from initial acquisition to final packaging. This structure ensures a comprehensive approach to data curation.

- **Stage 1: Data Acquisition**: Data will be collected from several web sources using a combination of REST APIs and web scraping techniques. Python scripts will be developed to handle API authentication, request throttling, and parsing of both JSON and HTML responses. All data provider terms of use will be respected and documented.
- **Stage 2: Quality Assessment and Cleaning**: Upon acquisition, raw data will be profiled to identify quality issues. This will involve checking for missing values (e.g., books without an ISBN), inconsistent formatting (e.g., *"J.K. Rowling"* vs. "*Joanne Rowling*"), and structural discrepancies between sources. A key task will be entity resolution: developing a robust process to match the same book across different datasets.
- **Stage 3: Data Integration and Modeling**: A unified data model will be designed to integrate the disparate sources. The International Standard Book Number (ISBN) will be used as the primary key where possible. Data will be transformed into a structured, analysis-ready format (likely a single CSV file or a relational SQLite database or both).
- **Stage 4: Workflow Automation and Provenance**: The entire curation process, from data fetching to final dataset generation, will be encapsulated in a single, automated workflow. This will likely be implemented as a master Python script. Snakemake and other relevant tools will also be explored. This ensures the process is transparent and can be re-run to update the dataset with new information.

- **Stage 5: Metadata and Documentation**: Comprehensive documentation will be created. This includes a data dictionary detailing each variable in the final dataset (description, data type, origin). A formal descriptive metadata will be created using the *schema.org* standard to ensure the dataset is discoverable and understandable.
- **Stage 6: Packaging and Dissemination**: The final project, including all scripts, the curated dataset, documentation, and environment specifications will be packaged in a GitHub repository to ensure it is understandable, reproducible, and reusable by others. Other ways to host will also be explored.

3. Data Sources

This project will integrate data from the following sources:

1. **Bibliographic Metadata**: The *Google Books API* will serve as the authoritative source for core metadata, including standardized titles, author names, publication dates, and, most importantly, ISBNs, which will be crucial for linking the other datasets.
2. **Literary Awards Data**: A list of winners and nominees for major English-language fiction awards (e.g., *Pulitzer Prize, National Book Award, Booker Prize*) from ***2000-2025*** will be scraped from their respective *Wikipedia* pages. Large Language Models will also be explored to see if they are more efficient in fetching and organizing data for cleaning.
3. **Reader Popularity Data**: Reader ratings and review counts will be acquired from the *Open Library API*. This source provides valuable metrics on public reception.
4. **Sales Performance Data**: Historical bestseller data will be fetched from *The New York Times Books API*. This will provide information on which books reached the bestseller list and for how many weeks.

   Kaggle Datasets: In case of unforeseen issues with APIs, datasets hosted on Kaggle like

Goodreads-books etc. will be explored for curating a combined dataset as per our requirements.

4. Team

This will be an individual project. All stages of the data curation lifecycle, from planning and acquisition to cleaning, documentation and final submission, will be completed by me.

5. Timeline

The project timeline would be attempted to align with the course topics covered and project milestones.

Weeks	Work	Target Date
2-3	Finalize data source selection and obtain necessary API keys. Develop initial exploratory scripts to understand the structure and limitations of each data source.	Sept. 29
4-5	Implement robust Python scripts for data acquisition from all four sources, including error handling and accommodating rate limits.	Oct. 13
6-7	Complete the first-pass integration of the datasets. Develop and apply initial data cleaning and entity resolution logic. A preliminary version of the merged dataset should be available.	Oct. 27 (Progress Report Due)
8-10	Refine the data cleaning and integration workflow. Begin drafting the data dictionary and creating the schema.org metadata record.	Nov. 17

11-12	Finalize the curated dataset. Automate the complete workflow. Write the narrative for the final project report.	Dec. 1
13	Package all project artifacts (code, data, documentation) into a GitHub repository. Conduct a final review and submit.	Dec. 10 (Final Submission Due)

6. Constraints

- **API Limitations**: All selected APIs have usage limits (rate limits, daily quotas). The acquisition scripts must be designed to respect these limits, which may slow down the data collection process. The availability and stability of these free APIs are also a potential risk to timelines.
- **Data Quality and Entity Resolution**: The primary technical challenge will be accurately matching books across disparate sources that may lack a common, clean identifier. While *ISBN* is the target, it may not be present in all records, necessitating fuzzy matching on title and author, which can be complex and error-prone.
- **Scope Management**: The initial scope is limited to three major awards and data from year *2000 onwards* to ensure the project is manageable within the semester timeframe.

7. Gaps

- **API Consistency**: The exact data structure and field availability within the APIs are not fully known and may contain unexpected inconsistencies or gaps that will only be discovered during implementation.

- **Web Scraping Fragility**: The structure of Wikipedia pages can change, which could break the web scraping scripts and require them to be updated. The plan must account for this potential maintenance.

8. References

1. Google Books APIs. (n.d.). Retrieved September 10, 2025, from https://developers.google.com/books
2. Open Library APIs. (n.d.). Retrieved September 10, 2025, from https://openlibrary.org/developers/api
3. The New York Times Developer Network. (n.d.). Books API. Retrieved September 10, 2025, from https://www.google.com/search?q=https://developer.nytimes.com/docs/books-product/1/overview.html