Workflow to Study Los Angeles's Media Presence through the Library of Congress Collection

Zhihui Zou[1]

[1]Duke University, Department of History and Department of Computational Media

Introduction

Los Angeles appears in our public memory accompanied by its Hollywood, real estate, and aerospace engineering industries. However, before them, the city's main industry was oil production. Edward Doheny drilled the first LA well in 1892, sparking an industry that once made LA one of the largest oil-producing regions in the world. In the 1920s, LA supplied a quarter of the world's oil. LA oil, a major financial, environmental, and supply agent, has hardly enjoyed as prominent of a media presence as its physical one in today's society. This project studies LA oil's existence in newspaper coverage between **1892 and 1930** to understand that LA's oil not only affected the US at large through providing fuel but also through various depictions in news media.

This project builds a workflow that allows users to access newspaper archives in the Library of Congress (LOC) easily at scale. Since LOC provides over 23 mil. newspaper pages, it could often be overwhelming for users. This study on LA's news appearance also acts as a test case for this workflow's functionalities to find areas for future academic and technical improvements.

Methodology

This in-progress project builds a workflow that:

1. Accesses the Library of Congress newspaper archive through the LOC API and entering specific search phrase;
2. Extracts newspaper pages and their publication metadata (e.g. publication city, page, etc.);
3. Store, clean, and manipulate the data in CSV, JSON, or other data formats;
4. Visually represents the data in Tableau or other data visualization methods.

The workflow's code is publically available on GitHub at: https://github.com/zzou21/LAOil

Workflow

Revising from LOC's API webpage (https://libraryofcongress.github.io/data-exploration/intro.html#historic-newspapers) and its provided code, this project traverses an LOC newspaper search result page in its JSON format in order to extract a search result's metadata.

```
# Accessing search URL:
searchURLTest = "https://www.loc.gov/newspapers/?end_date=1930-01-01&ops=~10&qs=los+angeles+oil&searchType=advanced&start_date=1892-04-20&
location_country=united+states&fo=json"
numberOfResults = 0
```

The search terms specified in the above query were:

Field: These words within 10 words of each other: "Los Angeles oil"
Date range: 04/20/1892 - 01/01/1930
Country: United States
Display level: Pages

The project extracts data through specifying which parts of a newspaper issue's metadata to extract. Here, we extract:

```
item_metadata_list.append({
    'Newspaper Title': Newspaper_Title,
    'Issue Date': Issue_Date,
    'Page Number': Page,
    'LCCN': LCCN,
    'City': City,
    'State': State,
    'Contributor': Contributor,
    'Batch': Batch,
    'PDF Link': pdf,
})
counter += 1
print(f"Processed {counter} results.")
```

Title of newspaper;
Issue date;
Page number;
Library of Congress Control Number;
City of newspaper publication;
State of publication;
Material contributor/contributing agency;
Archive batch;
Link to LOC PDF.

```
12810   Processed 12517 results.
12811   Processed 12518 results.
12812   Processed 12519 results.
12813   Processed 12520 results.
12814   Processed 12521 results.
12815
12816   Success! Ready to proceed to the next step!
12817   Finished compiling CSV
12818
```

Because my local machine does not have the memory to process all the search results at once (there are around 14,000 matching search results on LOC), I used Duke University's Duke Compute Cluster for extra computing power.

Data Manipulation Scenarios

```
,Newspaper Title,Issue Date,Page Number,LCCN,City,State,Contributor,Batch,PDF Link
0,['The Paducah sun.'],03-21-1901,3,['sn85052116'],['paducah'],['kentucky'],"['University of Kentucky, Lexington, KY']",
['kyu_liberace_ver01'],https://tile.loc.gov/storage-services/service/ndnp/kyu/batch_kyu_liberace_ver01/data/sn85052116/
00175044218/1901032102/0306.pdf
1,['The Paducah sun.'],04-23-1901,3,['sn85052116'],['paducah'],['kentucky'],"['University of Kentucky, Lexington, KY']",
['kyu_liberace_ver01'],https://tile.loc.gov/storage-services/service/ndnp/kyu/batch_kyu_liberace_ver01/data/sn85052116/
00175044218/1901042301/0480.pdf
```

Step 1: Raw data CSV

After accessing LOC's results using its API, the workflow stored raw data as a CSV. To turn this raw data into applicable geospatial data, a few data transformation and merging steps were performed.

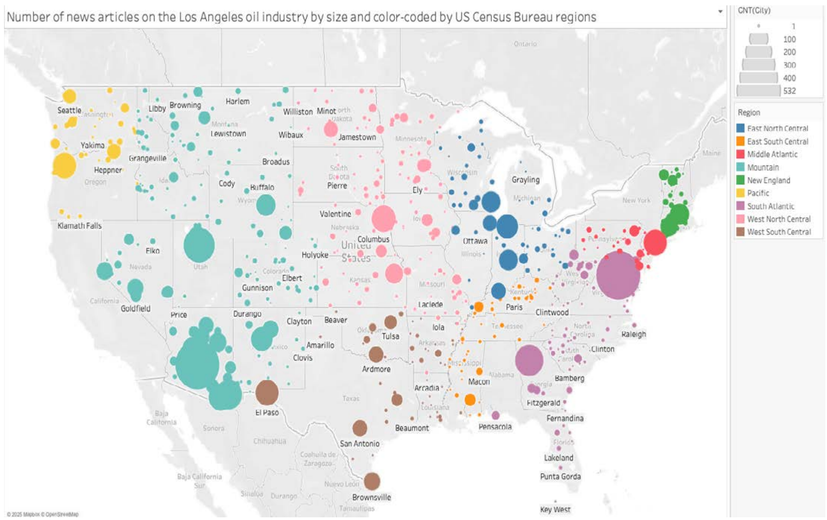
```
Unnamed: 0,Newspaper Title,Issue Date,Page Number,City,PDF Link
0,The Paducah Sun,03-21-1901,3,"Paducah, Kentucky",https://tile.loc.gov/storage-services/service/ndnp/kyu/
batch_kyu_liberace_ver01/data/sn85052116/00175044218/1901032102/0306.pdf
1,The Paducah Sun,04-23-1901,3,"Paducah, Kentucky",https://tile.loc.gov/storage-services/service/ndnp/kyu/
batch_kyu_liberace_ver01/data/sn85052116/00175044218/1901042301/0480.pdf
```

Step 2: Cleaned brackets & punctuations. Removed LCCN, Contributor & Batch columns. Standardized City & State columns to prepare for geographic plotting.

```
Newspaper Title,Issue Date,Page Number,City,State,PDF Link,prim_long_dec,prim_lat_dec,Region
The Paducah Sun,03-21-1901,3,Paducah,Kentucky,https://tile.loc.gov/storage-services/service/ndnp/kyu/batch_kyu_liberace_ver01/
data/sn85052116/00175044218/1901032102/0306.pdf,-88.6322692,37.0579342,East South Central
The Paducah Sun,04-23-1901,3,Paducah,Kentucky,https://tile.loc.gov/storage-services/service/ndnp/kyu/batch_kyu_liberace_ver01/
data/sn85052116/00175044218/1901042301/0480.pdf,-88.6322692,37.0579342,East South Central
```

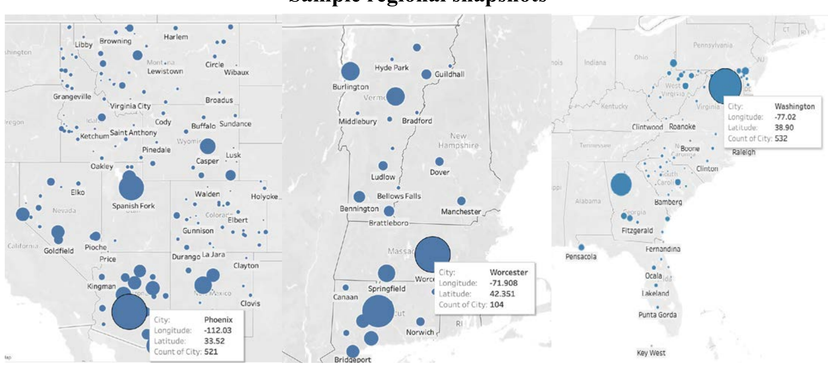
Step 3: Merged city coordinates and census region into the data frame. City coordinates were obtained from the Geographic Names Information System (GNIS). Region designations were obtained from the US Census Bureau's 9 census regions.

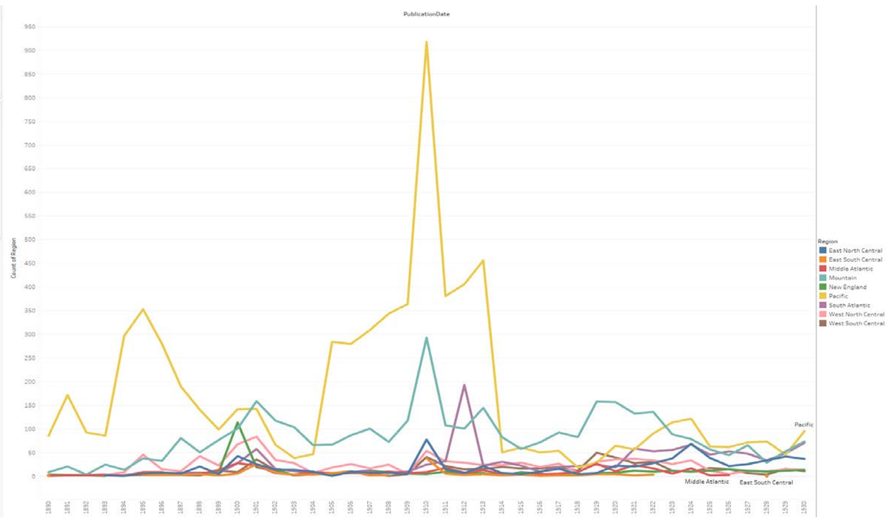
Visualization



Search results visualized by publication city and color-coded by census region (aggregated data with California excluded to avoid data skewing). Note that not all newspapers are included in LOC, so the visualization can only say so much about LA oil's presence in newspapers archived by LOC rather than the US at large.

Sample regional snapshots





Changes in number of search results by census region throughout the years. The yellow line is the Pacific region.

Workflow Analysis

This workflow allows users to access newspaper search results at scale without having to manually traverse LOC's search results. Data download has proven to be an easy process. However, data cleaning and manipulation required a comparatively significant amount of coding effort that, at this point, might or might not be easily transferred to a no-code UI. After data cleaning and manipulation have been completed, it is relatively convenient to integrate the cleaned data into visualization software like Tableau.

This workflow needs to consider if Tableau is the suitable workflow end-point due to both Tableau's cost and that integrating Tableau means limiting data processing to visualization and not any other possible pathways such as Natural Language Processing and more detailed geospatial analysis through ArcGIS or other tools.

Next Steps

The workflow has two future steps:
1) Implement a webpage UI to make this tool more accessible to users without requiring programming skills.
2) Implement the OCR texts that LOC provides for each newspaper page to conduct textual analysis alongside of geospatial analysis.

Acknowledgement

The inception of this project was done under the guidance of Hannah Jacobs and with the support from my peers in Hannah Jacobs's CMAC 310 class. I would also like to thank Trudi Abel, Mélanie Lamotte, and Daniel Ohayon for their time and support. Some parts of the code was modified from the template provided on the Library of Congress's API support website, and part of the computing process was done through using the Duke Compute Cluster.

Selected Bibliography

Mike Davis, *City of Quartz: Excavating the Future in Los Angeles* (New York: Verso Books, 1990).
Maxwell Johnson, "Borderlands Fortress: Newspaper Magnates, Preparedness, and the Rhetoric of Progress in World War I–Era Los Angeles," *Pacific Historical Review* 86, no. 2 (2017): 258-289.
Paul Sabin, *Crude Politics: The California Oil Market, 1900 - 1940* (Berkeley: University of California Press, 2005).
Lindsay Thomas and Abigail Droge, "The Humanities in Public: A Computational Analysis of US National and Campus Newspaper," *Journal of Cultural Analytics* 7, no. 1 (2022): 36-80.
Nancy Quam-Wickham, "'Cities Sacrificed on the Altar of Oil,' Popular Opposition to Oil Development in 1920s Los Angeles," *Environmental History* 3, no. 2 (1998): 189-209.