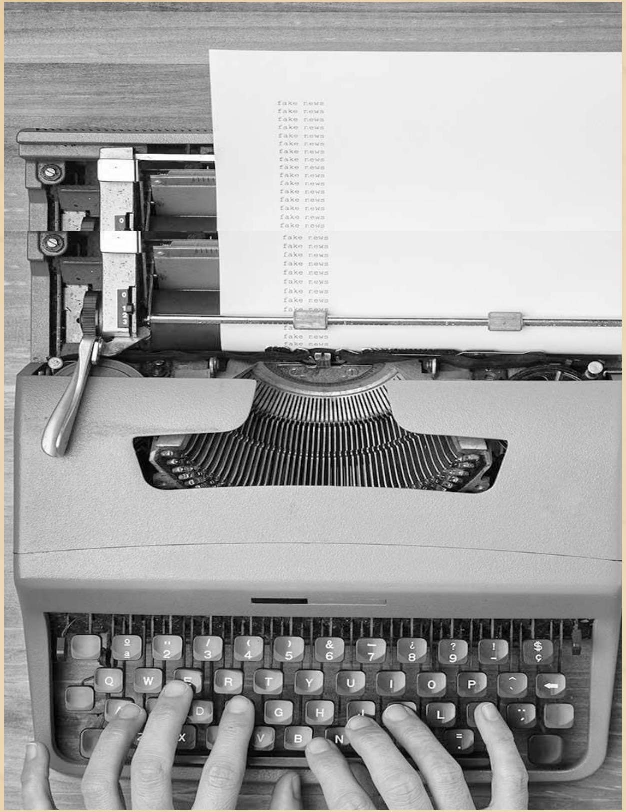


TEAM DATA WIZ FOR WASHINGTON DIGITAL NEWSPAPERS

Aayushi Gandhi, Manthan Mehta, Harshi Thaker

SUNDAY • JUNE 4 • 2023 **WASHINGTON STATE LIBRARY** EDITION № 001

EXECUTIVE SUMMARY

Overview.

The Washington Digital Newspapers (WDN) is an online archive of historically significant newspapers that can be used to produce data visualizations, apps for mobile devices, educational materials, or other data presentation techniques. In our capstone project, Team Data Wiz has created an interactive data visualization for the Washington State Library for the purpose of visualizing this archive in a more cohesive format. This in turn aids in answering any questions users might have about the archive.

Purpose.

The Washington State Library came to Team Data Wiz to find a way to better analyze their online archive of historically significant newspapers.

The Washington Digital Newspaper program has a repository of over 600,000 newspapers that are digitized over time (1852-2021). However, not all of this is stored in a single place for them to use for deriving insights. Team Data Wiz was tasked with creating a single dataset of the record of newspapers, including additional informative details about publication county, dates and number of issues to help perform data analysis on the same.

Objectives.

The main goal of our project is to lend structure to raw data in order to ultimately be able to help users at Washington State Library answer questions that would enable them to derive insights from the historic data.

In order to meet our project objectives, we followed a rigorous process of collaboration with our primary stakeholders, additional background research, data extraction, data wrangling and transformation. In doing so, Team Data Wiz, in consensus with our stakeholder at Washington State Library, reached upon a conclusion that an interactive data visualization is the best solution.

Result and Impact.

In our final deliverable, we have curated an interactive dashboard that consists of data visualizations that will answer all the logistic questions that users at Washington State Library have about the online archive of newspapers.

The strongest impact our solution has is shifting the tedious process into something that is streamlined and efficient. Moreover, it provides one single place where all the information about any digitized newspaper can be found, unlike their current state, which does not include the entire data set.

ORGANIZATION OVERVIEW

WASHINGTON
Secretary of State
Washington State Library

Summary and Background.

The Washington State Library is a government agency located in Tumwater, Washington, and serves as the library for state government and the reference library for all Washington residents. It is a part of the Office of the Secretary of State in Washington State.

The Washington State Library aims to provide access to information and resources for government agencies, state employees, and the general public. Its primary focus is to support the research, educational, and information needs of state government and libraries across Washington.

Washington Digital Newspapers.

Washington State Library houses a vast collection of digitized materials, including historical documents, photographs, newspapers, maps, and audiovisual materials. These collections are accessible online, providing valuable resources for researchers and the public.

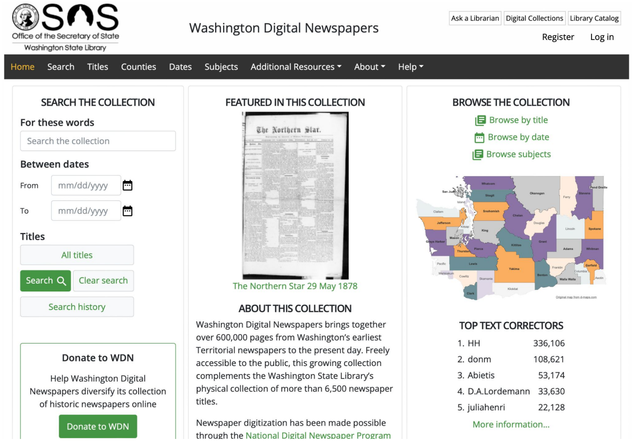
The Washington Digital Newspaper Program is an initiative undertaken by the Washington State Library to digitize and provide online access to historical newspapers from Washington State. The program aims to preserve and make available these valuable resources for researchers, historians, genealogists, and the general public.

Our Capstone project falls under the jurisdiction of the WDN program.

Stakeholders.

Shawn Schollmeyer, Washington Digital Newspapers Coordinator.

Shawn has been our primary stakeholder for the Capstone project.



Washington Digital Newspapers

SEARCH THE COLLECTION

For these words

Between dates

From

To

Titles

Donate to WDN

Help Washington Digital Newspapers diversify its collection of historic newspapers online

FEATURED IN THIS COLLECTION

The Northern Star 29 May 1878

ABOUT THIS COLLECTION

Washington Digital Newspapers brings together over 600,000 pages from Washington's earliest Territorial newspapers to the present day. Freely accessible to the public, this growing collection complements the Washington State Library's physical collection of more than 6,500 newspaper titles.

Newspaper digitization has been made possible through the National Digital Newspaper Program

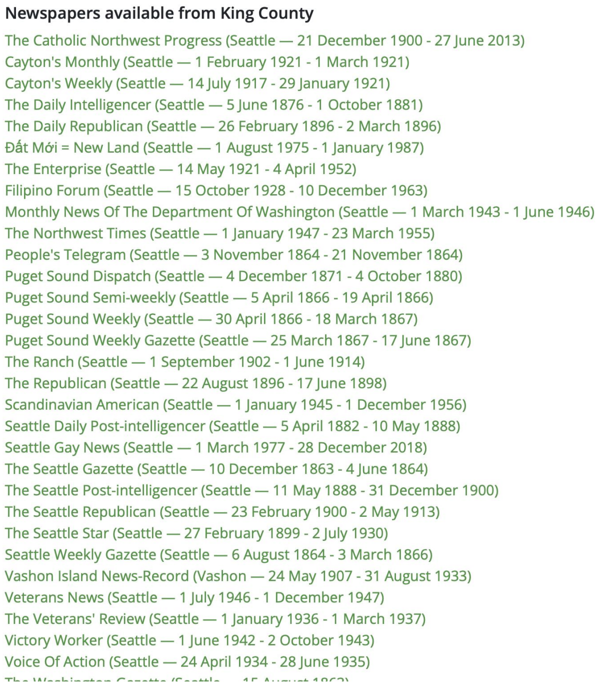
BROWSE THE COLLECTION

TOP TEXT CORRECTORS

1. HH	336,106
2. donm	108,621
3. Abietis	53,174
4. D.A.Lordemann	33,630
5. juliahenri	22,128

More information...

ISSUES AND OPPORTUNITIES



Newspapers available from King County

The Catholic Northwest Progress (Seattle — 21 December 1900 - 27 June 2013)
Cayton's Monthly (Seattle — 1 February 1921 - 1 March 1921)
Cayton's Weekly (Seattle — 14 July 1917 - 29 January 1921)
The Daily Intelligencer (Seattle — 5 June 1876 - 1 October 1881)
The Daily Republican (Seattle — 26 February 1896 - 2 March 1896)
Đất Mới = New Land (Seattle — 1 August 1975 - 1 January 1987)
The Enterprise (Seattle — 14 May 1921 - 4 April 1952)
Filipino Forum (Seattle — 15 October 1928 - 10 December 1963)
Monthly News Of The Department Of Washington (Seattle — 1 March 1943 - 1 June 1946)
The Northwest Times (Seattle — 1 January 1947 - 23 March 1955)
People's Telegram (Seattle — 3 November 1864 - 21 November 1864)
Puget Sound Dispatch (Seattle — 4 December 1871 - 4 October 1880)
Puget Sound Semi-weekly (Seattle — 5 April 1866 - 19 April 1866)
Puget Sound Weekly (Seattle — 30 April 1866 - 18 March 1867)
Puget Sound Weekly Gazette (Seattle — 25 March 1867 - 17 June 1867)
The Ranch (Seattle — 1 September 1902 - 1 June 1914)
The Republican (Seattle — 22 August 1896 - 17 June 1898)
Scandinavian American (Seattle — 1 January 1945 - 1 December 1956)
Seattle Daily Post-intelligencer (Seattle — 5 April 1882 - 10 May 1888)
Seattle Gay News (Seattle — 1 March 1977 - 28 December 2018)
The Seattle Gazette (Seattle — 10 December 1863 - 4 June 1864)
The Seattle Post-intelligencer (Seattle — 11 May 1888 - 31 December 1900)
The Seattle Republican (Seattle — 23 February 1900 - 2 May 1913)
The Seattle Star (Seattle — 27 February 1899 - 2 July 1930)
Seattle Weekly Gazette (Seattle — 6 August 1864 - 3 March 1866)
Vashon Island News-Record (Vashon — 24 May 1907 - 31 August 1933)
Veterans News (Seattle — 1 July 1946 - 1 December 1947)
The Veterans' Review (Seattle — 1 January 1936 - 1 March 1937)
Victory Worker (Seattle — 1 June 1942 - 2 October 1943)
Voice Of Action (Seattle — 24 April 1934 - 28 June 1935)
The Washington Gazette (Seattle — 15 August 1863)

Issues.

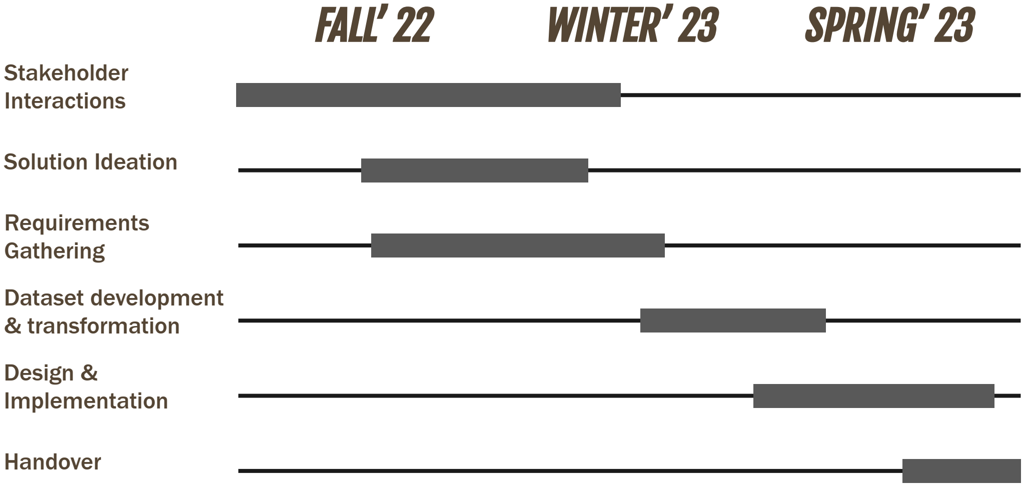
The current state of the WDN program is as depicted in the image alongside - the data is unstructured. This brings up the problem of overlooking important information. There may also be a scenario where it can get difficult to completely understand where the gaps lie in the current data that is available for analysis. This again leads to poor data analysis and derivation of insights from the data.

Opportunities.

Our list of issues presents us with opportunities to work with. We came up with a list of things we want improve for by means of our Capstone project. Firstly, we made a brand new dataset by pulling data from multiple sources to establish one source of truth. Secondly, we used Tableau to create interactive dashboards that lended structure to the data available. Thirdly, we published this onto Tableau Public which gives our work visibility not just within the state of Washington, bust also all around the globe. Lastly, we coached our primary stakeholder on the maintenance of the interactive dashboards enabling them to keep updating it as and when they are presented with new data.

APPROACH

1 - Project Schedule plan



- The project was approached in a structured manner with a major focus on solving for the right problem using an optimized project management and agile approach.
- A schedule along with Key performance indicators (KPI) was set up during initial stakeholder discussions to ensure there were defined deliverables, deadlines, action items and regular check-ins and project updates.
- This ensured the project was on track, adhere to the deadlines, escalate issues and eliminate any unforeseen challenges

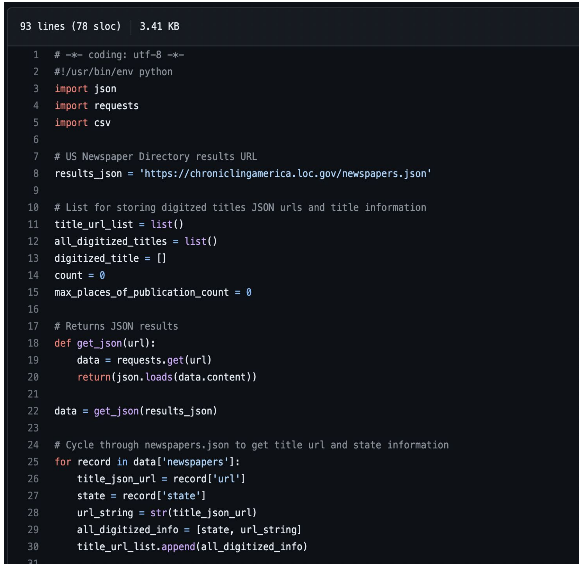
APPROACH

2 - Data Extraction

In order to cater the needs and requirements of the stakeholder, i.e. perform data analysis, we need to first get ahold of the data available.

All data related to the digitized newspapers was made available to the public via the Chronicling America website and one could navigate the website to find information related to any newspaper that was digitized.

The issue was that Washington state library did not have a well-defined or structured database as it was storing all the information regarding the digitized newspapers on the Chronicling America Website. Hence, the first step in this process was to extract all the data available on the website and store it in a place where it can be cleaned & transformed for further analysis.



93 lines (78 sloc) 3.41 KB

```python
# -*- coding: utf-8 -*-
#!/usr/bin/env python
import json
import requests
import csv

# US Newspaper Directory results URL
results_json = 'https://chroniclingamerica.loc.gov/newspapers.json'

# List for storing digitzed titles JSON urls and title information
title_url_list = list()
all_digitized_titles = list()
digitized_title = []
count = 0
max_places_of_publication_count = 0

# Returns JSON results
def get_json(url):
    data = requests.get(url)
    return(json.loads(data.content))

data = get_json(results_json)

# Cycle through newspapers.json to get title url and state information
for record in data['newspapers']:
    title_json_url = record['url']
    state = record['state']
    url_string = str(title_json_url)
    all_digitized_info = [state, url_string]
    title_url_list.append(all_digitized_info)
```

Data Extraction using Python

To extract the data from the Chronicling America website, we developed a Python script that interacts with the Chronicling America API to collect basic metadata about all available digitized newspapers into a csv file. The metadata retrieved in this step includes the newspaper names, publication locations, dates, and quantities of digitized issues available in the system.

Python's simplicity, powerful libraries, extensive community support, and flexibility made it an excellent choice for web scraping.

Additional data was brought in from a third-party vendor directly in json format and specific data was extracted, resulting in a comprehensive data set.

APPROACH

3 - Data Wrangling & Transformation

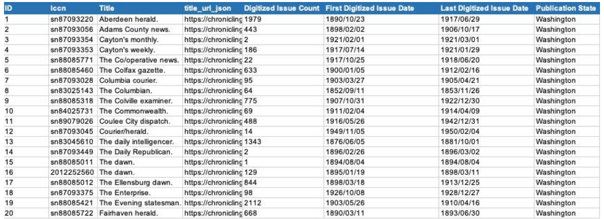
The raw data was collected and accumulated into a csv file. Based on the Stakeholder discussions, the resulting analysis was to be done based in data specific to the counties in the Washington State. Taking this into consideration, and also the fact that the resulting data set was small about 113 records, we decided to use Excel for data wrangling and transformation purposes.

Data Wrangling using Excel

The data captured from the website was unstructured and not fit for analysis. Some Data Wrangling was performed to make the data consistent and structured.

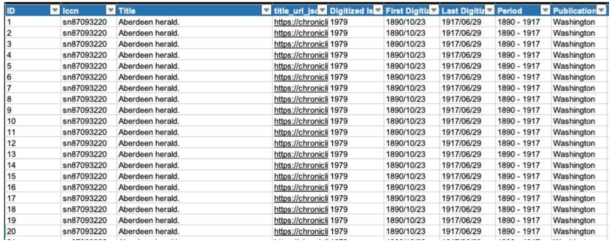
- Dates were formatted into a yy/mm/dd format
- Unnecessary columns from the dataset were eliminated/removed

Original Scraped data set (113 rows)



ID	lccn	Title	title_url_json	Digitized Issue Count	First Digitized Issue Date	Last Digitized Issue Date	Publication State
1	sn87093220	Aberdeen herald.	https://chroniclinc	1979	1890/10/23	1917/06/29	Washington
2	sn87093056	Adams County news.	https://chroniclinc	443	1898/02/02	1906/10/17	Washington
3	sn87093354	Cayton's monthly.	https://chroniclinc	2	1921/02/01	1921/03/01	Washington
4	sn87093353	Cayton's weekly.	https://chroniclinc	186	1917/07/14	1921/01/29	Washington
5	sn88085771	The Co/operative news.	https://chroniclinc	22	1917/10/25	1918/06/20	Washington
6	sn88085460	The Colfax gazette.	https://chroniclinc	633	1900/01/05	1912/02/16	Washington
7	sn87093028	Columbia courier.	https://chroniclinc	95	1903/03/27	1905/04/21	Washington
8	sn83025143	The Columbian.	https://chroniclinc	64	1852/09/11	1853/11/26	Washington
9	sn88085318	The Colville examiner.	https://chroniclinc	775	1907/10/31	1922/12/30	Washington
10	sn84025731	The Commonwealth.	https://chroniclinc	69	1911/02/04	1914/04/09	Washington
11	sn89079026	Coulee City dispatch.	https://chroniclinc	488	1916/05/26	1942/12/31	Washington
12	sn87093045	Courier/herald.	https://chroniclinc	14	1949/11/05	1950/02/04	Washington
13	sn83045610	The daily intelligencer.	https://chroniclinc	1343	1876/06/05	1881/10/01	Washington
14	sn87093449	The Daily Republican.	https://chroniclinc	2	1896/02/26	1896/03/02	Washington
15	sn88085011	The dawn.	https://chroniclinc	1	1894/08/04	1894/08/04	Washington
16	2012252560	The dawn.	https://chroniclinc	129	1895/01/19	1898/03/11	Washington
17	sn88085012	The Ellensburg dawn.	https://chroniclinc	844	1898/03/18	1913/12/25	Washington
18	sn87093375	The Enterprise.	https://chroniclinc	98	1926/10/08	1928/12/27	Washington
19	sn88085421	The Evening statesman.	https://chroniclinc	2112	1903/05/26	1910/04/16	Washington
20	sn88085722	Fairhaven herald.	https://chroniclinc	668	1890/03/11	1893/06/30	Washington

Drilled Down data set (84k rows)



ID	lccn	Title	title_url_jsc	Digitized Is	First Digitiz	Last Digitiz	Period	Publication
1	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
2	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
3	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
4	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
5	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
6	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
7	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
8	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
9	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
10	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
11	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
12	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
13	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
14	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
15	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
16	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
17	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
18	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
19	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington
20	sn87093220	Aberdeen herald.	https://chronicli	1979	1890/10/23	1917/06/29	1890 - 1917	Washington

- New columns such as ID, Period, Issued Date, subject & language were included

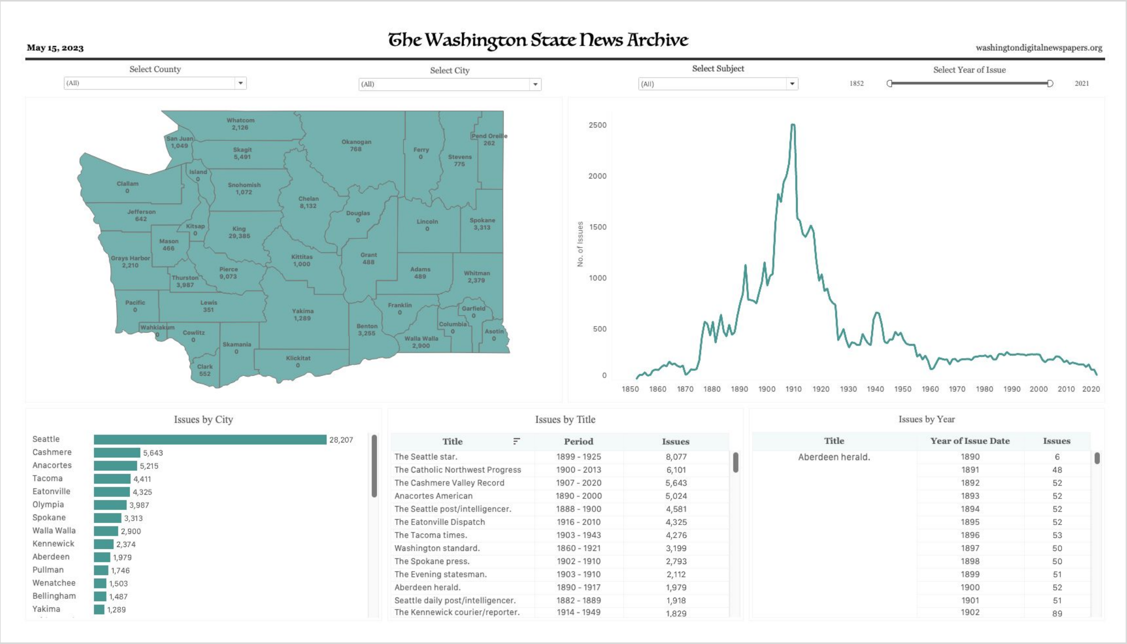
Additional Data Acquisition from JSON files

The summarized data set consisted of around 113 records with 113 different newspaper titles in Washington State. However, to provide a time series analysis, the data needed to be drilled down for each newspaper title with specific information about all the dates when the newspaper was available. This data was manually extracted using the json urls extracted from the website resulting in an enhanced data set of 84k rows

An online tool was used to convert the json urls into csv files and then the digitized dates columns were created.

DELIVERABLES

Digitized Newspaper Data Visualization



May 15, 2023

The Washington State News Archive

washingtondigitalnewspapers.org

Select County: (All)

Select City: (All)

Select Subject: (All)

Select Year of Issue: 1852 – 2021

Issues by City

Seattle	28,207
Cashmere	5,643
Anacortes	5,215
Tacoma	4,411
Eatonville	4,325
Olympia	3,987
Spokane	3,313
Walla Walla	2,900
Kennewick	2,374
Aberdeen	1,979
Pullman	1,746
Wenatchee	1,503
Bellingham	1,487
Yakima	1,289

Issues by Title

Title	Period	Issues
The Seattle star.	1899 - 1925	8,077
The Catholic Northwest Progress	1900 - 2013	6,101
The Cashmere Valley Record	1907 - 2020	5,643
Anacortes American	1890 - 2000	5,024
The Seattle post/intelligencer.	1888 - 1900	4,581
The Eatonville Dispatch	1916 - 2010	4,325
The Tacoma times.	1903 - 1943	4,276
Washington standard.	1860 - 1921	3,199
The Spokane press.	1902 - 1910	2,793
The Evening statesman.	1903 - 1910	2,112
Aberdeen herald.	1890 - 1917	1,979
Seattle daily post/intelligencer.	1882 - 1889	1,918
The Kennewick courier/reporter.	1914 - 1949	1,829

Issues by Year

Title	Year of Issue Date	Issues
Aberdeen herald.	1890	6
	1891	48
	1892	52
	1893	52
	1894	52
	1895	52
	1896	53
	1897	50
	1898	50
	1899	51
	1900	52
	1901	51
	1902	89

The final deliverable is a **Tableau data visualization dashboard** which

- Provides a dynamic and interactive approach to understanding the current data
- Helps in identifying gaps in the current data
- Serves as a platform to market the data as well as increase reach

Highlights

- Filters gives the power to focus on key areas of interest
- Insights into newspaper digitized by county, city, period
- Temporal insights of digitized newspaper issues
- Geographical view of newspaper issues by counties & cities - additional spatial file was brought in & joined with the csv file to create the map

BENEFITS

The Ultimate Platform

The project brings together all the identified data, county representation, and temporal information onto a single dashboard. This means that users can access and view all the relevant information in one place, simplifying the process of data exploration and analysis.The project also provides insights into each county's representation within the digitized issues. It allows users to visualize and analyze the data over time, offering a temporal perspective on the digitized content.

Bridging Gaps in Data

WSL has a repository of all of the digitized issues that they have on their website, but in an unclean format, which makes it difficult to navigate through, or identify where there are any missing issues. Through visualization on the dashboard, users can identify gaps or missing information within the digitized issues. This helps users recognize areas where information is incomplete and prompts further investigation.

Widen the Reach

The project enables WSL to identify all the data they have available, specifically digitized issues that can be accessed on their website. This means that the project helps in recognizing and organizing the existing digital content. The dashboard created as part of the project emphasizes ease-of-use. Its intuitive design and functionality make it an effective tool for displaying all the information in a centralized manner, achieving the goal of providing a user-friendly experience for data exploration.

Interactive Insights

The project leverages Tableau Public, a data visualization tool, to make the information easily accessible to users worldwide. By utilizing Tableau Public, the project team has succeeded in cleaning, filtering, and presenting the current information in a user-friendly manner. The dashboard allows users to filter and customize the data according to their preferences. This high level of interactivity empowers users to explore the data and extract insights based on their specific needs.

IMPACT & NEXT STEPS

Impact

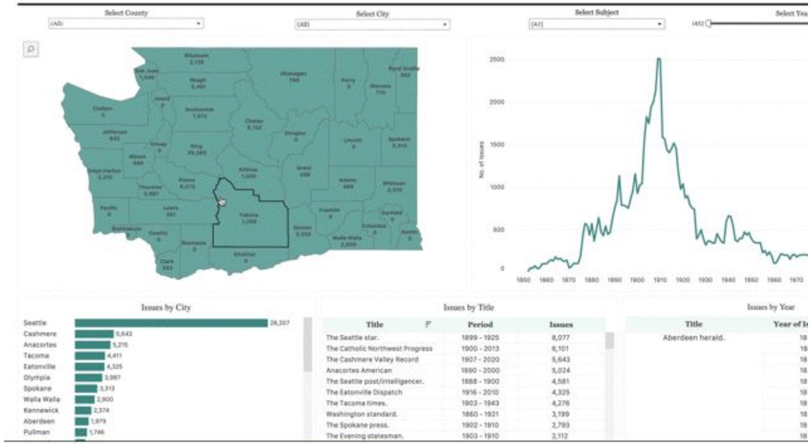
- Data Organization: The project organizes all available digitized issues for easy access for now as well as the future.

- Insights and Analysis: Users gain valuable insights and can analyze trends and patterns in the data.

- Missing Information Identification: The project helps identify gaps in the digitized data for further investigation.

- Global Accessibility: Integration with Tableau Public allows worldwide access to the data for the long term.

- User-Friendly Interface: The user-friendly dashboard enables easy exploration and filtering of the data.



Issues by Title

Title	Period	Issues
The Seattle star.	1899 - 1925	8,077
The Catholic Northwest Progress	1900 - 2013	6,101
The Cashmere Valley Record	1907 - 2020	5,643
Anacortes American	1890 - 2000	5,024
The Seattle post/intelligencer.	1888 - 1900	4,581
The Eatonville Dispatch	1916 - 2010	4,325
The Tacoma times.	1903 - 1943	4,276
Washington standard.	1860 - 1921	3,199
The Spokane press.	1902 - 1910	2,793
The Evening statesman.	1903 - 1910	2,112

Issues by Year

Title	Year of Is
Aberdeen herald.	18
	18
	18
	18
	18
	18
	18
	18
	18
	18

Next Steps

- Hosting on the Website: The final product, which includes the user-friendly dashboard and the digitized data, needs to be deployed on the WSL website.
- Integration with Tableau Public: The WSL team needs to integrate the dashboard with the Tableau Public profile of WSL. This integration will allow users to access the dashboard directly from the Tableau Public page of WSL.
- Continuous Maintenance and Updates: Ongoing maintenance is required to ensure the dashboard remains functional and up to date. This includes regular monitoring of data sources, addressing any technical issues or bugs, and incorporating user feedback for improvements.

LESSONS LEARNED

1. **Scope Management**: Defining the project scope and managing its boundaries can be challenging. It's common to have ambitious goals that may not be feasible within the project's timeframe or resources. However, it was crucial to have set boundaries before initiating a project.
   Approach: *Clearly defining project objectives and scope from the beginning is essential. Regularly revisiting and refining the scope as the project progresses helps in maintaining focus and managing expectations.*

2. **Data Availability and Quality:** Obtaining relevant and reliable data can be a significant challenge. It may require thorough research, contacting data providers, or dealing with incomplete or messy datasets. Exploratory data analysis is a crucial step in any data centric project.
   Approach: *Planning ahead and allowing ample time for data acquisition and cleaning was crucial. It's important to explore available data sources and have backup plans in case of data limitations.*

3. **Time Management:** Balancing the project's workload with other academic or personal commitments can be a challenge. Meeting deadlines, allocating sufficient time for each project phase, and managing unexpected delays can be demanding.
   Approach: *To overcome the same & ensure effective time management requires setting realistic milestones, and maintaining regular progress tracking. Breaking down the project into smaller tasks with deadlines helps in staying on track.*

4. **Communication and Collaboration:** Collaborating with teammates, advisors, or stakeholders may involve challenges such as miscommunication, conflicting opinions, or difficulty in coordinating schedules.
   Approach: *Establishing clear communication channels, setting regular meetings, and fostering an open and respectful environment for discussion and feedback are essential. Active listening & adapting communication styles to different stakeholders facilitate effective collaboration.*

5. **Technical Challenges:** Working with complex data analysis tools, and programming languages can present technical hurdles. Debugging errors, handling computational limitations, or implementing advanced techniques may require additional learning and troubleshooting.
   Approach: *Maintaining a learning mindset and seeking help from mentors or online communities can overcome technical challenges. Breaking down complex tasks into smaller, manageable steps also aids in tackling technical difficulties.*

REFERENCES

[1] National Endowment for the Humanities. (n.d.). *Chronicling America: Library of Congress*. Chronicling America API: https://chroniclingamerica.loc.gov/about/api/

[2] Ferriter, M. (2019, May 21). Visualizing Chronicling America Data: 15 million pages of digitized historical newspapers: The Signal. The Library of Congress. https://blogs.loc.gov/thesignal/2019/05/visualizing-chronicling-america-data-15-million-pages-of-digitized-historical-newspapers/

[3] Paranick, A. (2022, July 14). New Interactive Map and Timeline added to Chronicling America: Headlines and heroes. The Library of Congress. https://blogs.loc.gov/headlinesandheroes/2022/07/new-chronicling-america-interactive-map-and-timeline/

[4] Veridian. (n.d.). Washington Digital Newspapers. https://washingtondigitalnewspapers.org/

[5] Washington State Geospatial Open Data Portal. (n.d.). https://geo.wa.gov/