Analyzing the Historic Maryland Newspapers Project Collection

PAMELA A. MCCLANAHAN
University of Maryland Libraries

Abstract: The Historic Maryland Newspapers Project (HMNP) is a mass digitization project to provide Maryland newspaper content to the National Digital Newspaper Program through the Chronicling America newspaper database at the Library of Congress. These digitized newspapers form an artificial digital collection of Maryland newspaper content. This is a case study that analyzes existing data about the Maryland digital newspaper collection, with two goals: to provide better understanding of an artificial collection for researchers, and to use data analysis to aid in selection decisions to prioritize diverse content for a large digital collection.

Introduction

The Historic Maryland Newspapers Project (HMNP) at University of Maryland Libraries is the Maryland state awardee of the National Digital Newspaper Program (NDNP), a partnership between Library of Congress (LC) and National Endowment for the Humanities (NEH). The NDNP program has created the Chronicling America online database with historic newspapers from nearly every state, which is openly available on the LC website.[1] NDNP awardees digitize approximately 100,000 newspaper pages every two years that are added to Chronicling America, which currently contains over 17 million pages.

NDNP was developed from the earlier U.S. Newspaper Program, a similar partnership between NEH and LC that awarded grants to a partner organization in every state to locate, catalog, and microfilm newspaper titles from across the state. USNP resulted in the U.S. Newspaper Directory, which is available to search on the Chronicling America website, and many of those microfilmed titles are now being digitized through NDNP.[2] Maryland's first newspaper was *The Maryland Gazette* published in Annapolis in 1727. However, Baltimore developed into the larger city and produced by far the most newspapers in the state and most surviving newspapers were published there. The Maryland State Archives served as the microfilm repository for the USNP in Maryland and holds most of the state newspaper collection now.

The Historic Maryland Newspapers Project (HMNP) has been participating in NDNP since 2012 and has digitized 316,168 newspaper pages.[3] HMNP is currently completing its fourth NDNP grant phase.[4] The extended grant goes through December 31, 2020 (NEH approved an extension due to the coronavirus pandemic). This phase will add another 100,000 pages of Maryland newspaper content to Chronicling America. Additionally, HMNP has been awarded a phase five NDNP grant from NEH, which runs

[1] "Chronicling America," Chronicling America: Historic American Newspapers, Library of Congress, accessed November 7, 2020, https://chroniclingamerica.loc.gov/.

[2] "U.S. Newspaper Directory, 1690-Present," Chronicling America: Historic American Newspapers, Library of Congress, accessed November 7, 2020, https://chroniclingamerica.loc.gov/search/titles/.

[3] "The Historic Maryland Newspapers Project," University Libraries, University of Maryland, accessed November 7, 2020, https://www.lib.umd.edu/digital/newspapers/home.

[4] "National Digital Newspaper Program," Division of Preservation and Access, National Endowment for the Humanities, accessed November 7, 2020, https://www.neh.gov/grants/preservation/ national-digital-newspaper-program.

Published by Society of American Archivists, September 2020.

September 1, 2020 to August 31, 2022. This fall, HMNP is working to finalize title selection from the identified title possibilities included in this next phase's grant application.

This study examines the title metadata from the digitized Maryland newspaper content to complete a bibliographic and geographic analysis of the collection of historic Maryland newspapers available in Chronicling America. The goal is to give researchers a better understanding of the Maryland content in Chronicling America and aid the selection process for the phase five NDNP grant.

Problem Statement

Varying by state, the NDNP newspaper project often represents a portion of the state awardee's newspaper holdings. In Maryland, neither HMNP nor UMD Libraries as a whole hold a physical print or microfilm 'historic Maryland newspaper collection.' HMNP as represented in Chronicling America is an artificial digital collection made up of newspapers published in Maryland, and the physical content is held by partner institutions throughout the state. HMNP coordinates loans of materials and manages the digitization of content, while adhering to NDNP technical standards for inclusion in Chronicling America. Previously a researcher would have needed to visit multiple repositories in the state to view these newspaper titles on microfilm or the original print, but now this content is collected and available in one online location. However since no single institution holds the original material, there is not an overall institutional finding aid or pathfinder specific to the content in Chronicling America and this 'collection' has not been studied as a whole entity.

This project examines the digital Maryland newspaper collection as a whole, in order to provide a better understanding of the historic Maryland newspapers in Chronicling America for researchers. In addition, this analysis has been conducted with the purpose of using this new understanding of the current collection to fill in gaps in current title offerings for a fifth grant phase. One of the main research questions for the project is the geographic distribution of content across the state including number of titles, number of issues, and number of pages for comparison.

Since this research project is intended to help with future title selection decisions, the NDNP selection criteria must be considered. Current NDNP selection criteria include:

- content must be published in state between 1690-1963 and be in the public domain;
- significant research value (papers of record for the state and good geographic, temporal, social, and political coverage of the state, as well as prioritizing content from diverse and underrepresented groups);
- good image quality, preferably from the master negative microfilm (but can consider positive microfilm and print);
- with as complete a run as possible - long runs, continuity, few to no gaps; and
- has not already been digitized elsewhere (unless there is a compelling reason to do so).[5]

NDNP also requires each state to have an advisory board to assist in the title selection process. The HMNP advisory board is made up of archivists, librarians, historians, educators, and other experts from across the state. The HMNP advisory board has prioritized titles from communities traditionally underrepresented in archival records including immigrant communities, women, political minority groups, and labor groups. Newspapers are often one of the few primary sources for these communities' activities as they are not always otherwise well-documented in the historical record. While the advisory board has brought forth some excellent titles and can provide anecdotal evidence for particular titles or

[5] "Content Selection," National Digital Newspaper Program, Library of Congress, accessed November 7, 2020, https://www.loc.gov/ndnp/guidelines/selection.html.

regions of the state to have more representation in the collection based on what is observed in their work, this research project aims to further analyze the existing title data to aid in the selection decision as well. This project will gather and evaluate data from the Chronicling America Maryland newspaper collection comparing the number of titles, issues, and pages with such factors as geographic distribution, and distribution of newspapers from underrepresented communities. The goal is to better understand the Maryland newspaper collection in Chronicling America to make future digitization selection decisions prioritizing diversity and to share this information with researchers and archives, libraries, and other cultural heritage institutions throughout the state.

A note on existing research: Another state newspaper project, the Georgia NDNP project has published an article on selecting newspaper titles for digitization, but the focus was on titles beyond NDNP for their larger digital library newspaper initiative. The article discusses many of the same selection criteria, but other criteria for non-NDNP projects as well. It also does not involve a data analysis of newspapers that have already been digitized.[6] In terms of data analysis, the Library of Congress provides data on the batches of images and metadata that has been submitted by states for inclusion in Chronicling America, but not down to the title level for all metadata. A batch may have multiple unrelated titles or a title may extend across multiple batches, which makes title level data more useful for this kind of analysis rather than batch level. Additionally, the Library of Congress often provides data and analysis for the program as a whole or occasionally to the state level.[7] This research specifically looks at title data for the state of Maryland's historic newspapers in Chronicling America and will go to the regional and county level beyond what has been made available by NDNP at the Library of Congress.

Methodology

This is a case study of the Historic Maryland Newspapers Project (HMNP) using title data from the digitized Maryland newspaper content in Chronicling America to complete a data analysis of this artificial digital collection that could be helpful for researchers and assist HMNP in future title selection decisions.

Title data was gathered from Chronicling America, the U.S. Newspaper Directory, and HMNP's own metadata collation of newspaper titles and input into an Excel spreadsheet.[8] Fields collected include Title, Date Range, City of Publication, County, Pages, Issues, and Language. Additional fields are Region and Underrepresented groups. These last two fields were manually determined and added by the Digital Projects Librarian. Underrepresented titles for the purposes of this project include non-English languages, immigrant communities, women, political minorities, and labor groups. These titles were intended specifically for these minority audiences in contrast to the paper of record for the town intended for the general population, or in most cases during this time period written by white men for white men.

Excel pivot tables were used to sum data including number of titles, number of issues, and number of pages by county and by region. Sorting and filters in Excel allowed for quick display of this information. Tableau Desktop was used to create map-based data visualizations based on guidance from a Storybench

[6] Donnie Summerlin, "Selecting Newspaper Titles for Digitization at the Digital Library of Georgia," *D-Lib Magazine* 20, nos. 9–10 (September–October 2014), https://doi.org/10.1045/september2014-summerlin.

[7] "Chronicling America Data Visualizations," National Digital Newspaper Program, Library of Congress, accessed November 7, 2020, https://www.loc.gov/ndnp/data-visualizations.html.

[8] Special acknowledgement to Bryanna Bauer (HMNP Student Assistant and MLIS candidate at UMD iSchool) for assisting with the data entry.

article[9] and Tableau Help.[10] Using Federal FIPS county codes[11] and ESRI US county shape data[12] along with the Excel pivot tables in Tableau, the authors created a map of the different sum factors by county with a density measure (lighter color = lower number and darker color = higher number). Four maps were created including Sum of Titles by County, Sum of Issues by County, Sum of Pages by County, and Sum of Underrepresented Group Titles by County. The maps are interactive allowing a user to hover over a polygon in the map to display the county name and the sum factor. Regional sums were not used at this time, but are easily visualized when seeing where a county is on the map and the density colors for the surrounding area.

After the SAA Research Forum, the language of the newspaper pages was analyzed as well to see how many of the total pages were English language vs. non-English language. This data is displayed in an Excel pie chart. Currently, all non-English language titles included in Chronicling America from Maryland were published in Baltimore City making a geographic analysis unnecessary.

Results

While a small state, Maryland has distinct regions with the Chesapeake Bay cutting through the state with eastern and western shores with their unique characteristics, as well as a mountainous western region and the more urban area of central and northern Maryland. Going into this research project one known result was that Baltimore City would have the most titles, issues, and pages by far compared to anywhere else in the state. Baltimore City was not only the largest city in Maryland, but one of the top five largest urban areas in the country through 1860 and one of the top ten until 1980.[13] Not including Baltimore City, Central Maryland counties still have the most newspaper pages. It is much more dispersed in Western Maryland and the Eastern Shore following demographic trends of the state.

Map 1 is a density map showing the sum of titles of underrepresented groups by county. Baltimore City has the highest number of titles coming from underrepresented groups with the ethnic press from the many diverse immigrant communities that settled in Baltimore. Maps 2 - 4 are density maps showing the sum of titles, issues, and pages by county. Map 5 is an interactive map that when the user hovers over a county a pop-up displays three sum factors (issues, titles, and pages in Chronicling America) for that county. This screenshot shows the example of hovering over Prince George's County, Maryland. Chart 1 is a pie chart comparing English language pages (68% of total pages) and non-English language pages (32% of total pages).

[9] Hanyang Dong, "How to build an interactive county level map in tableau," Storybench, School of Journalism, Northeastern University, accessed November 7, 2020, https://www.storybench.org/build-interactive-county -level-map-tableau/.

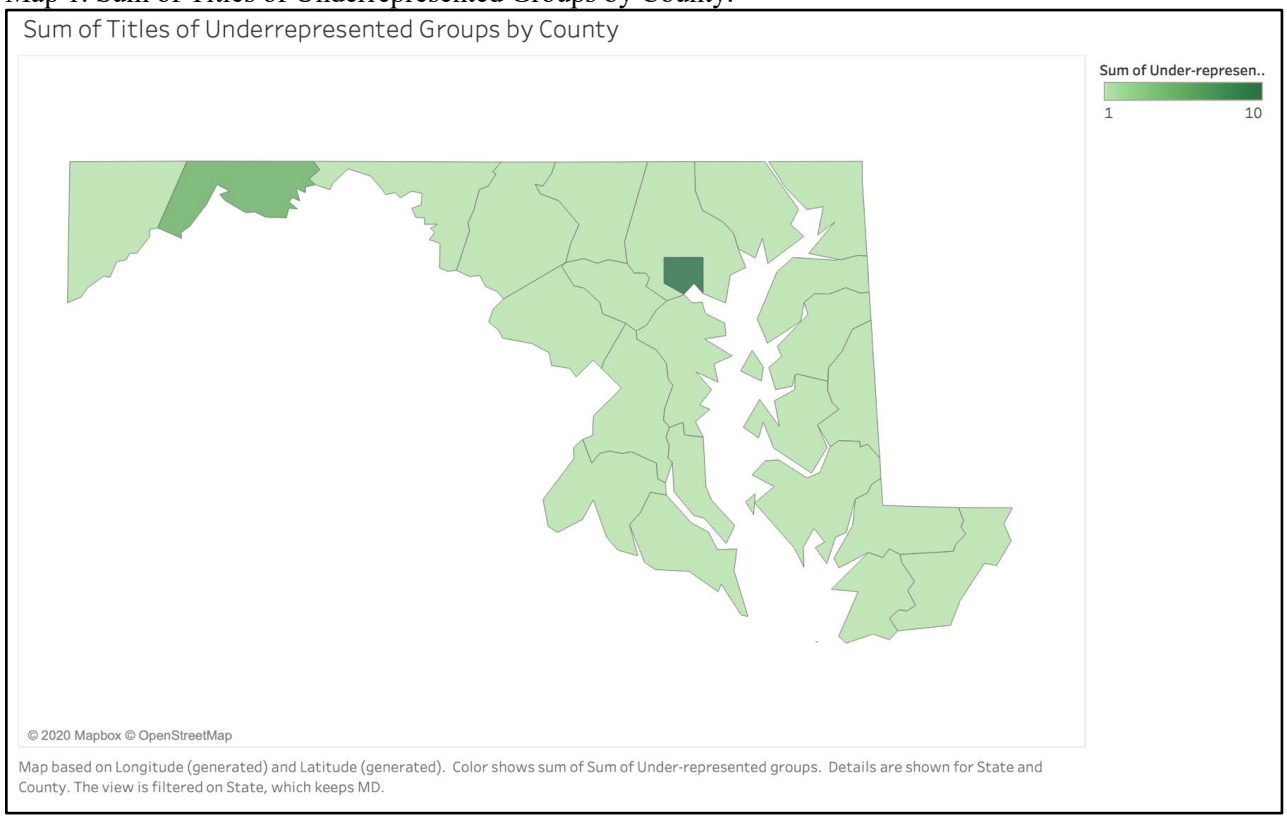
[10] "Maps and Geographic Data Analysis in Tableau," Tableau Help, accessed November 7, 2020, https://help.tableau.com/current/pro/desktop/en-us/maps.htm.

[11] "County FIPS Codes," Natural Resources Conservation Service, U.S. Department of Agriculture, accessed November 7, 2020, https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697.

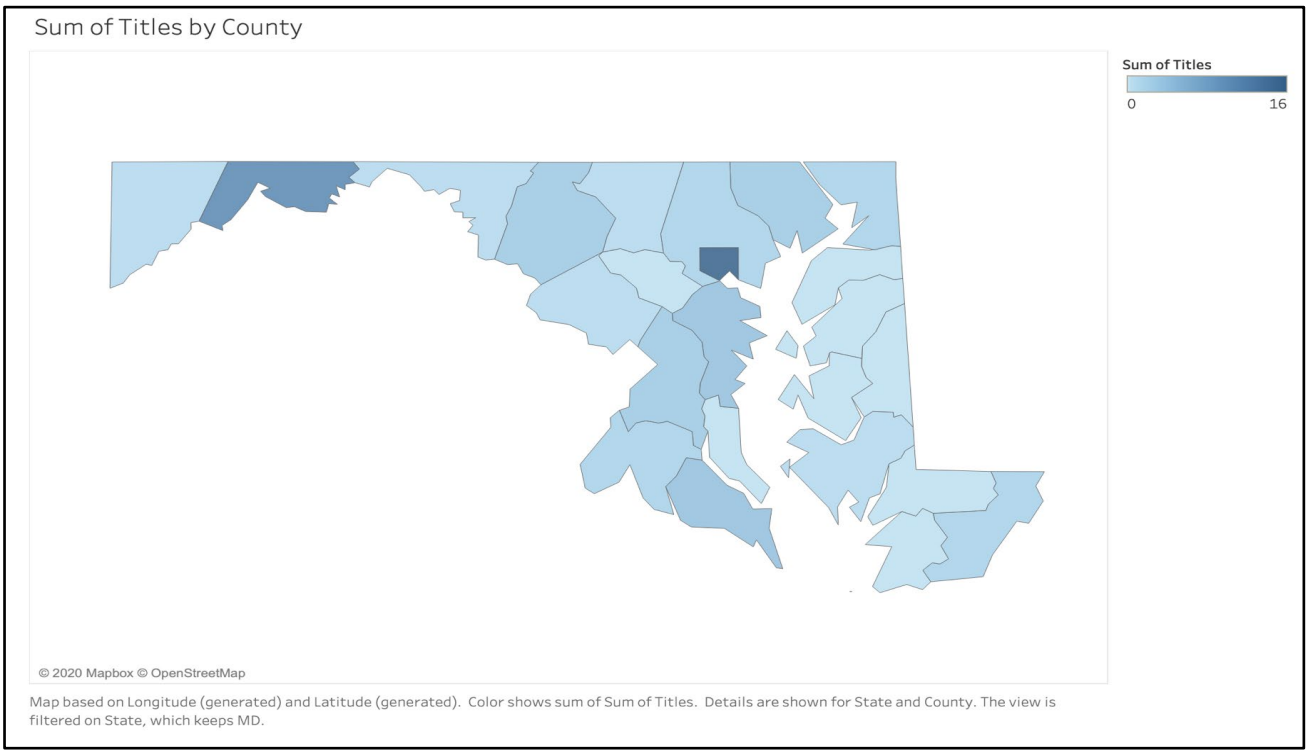
[12] "Uscounties.shp," Community, Esri, accessed November 7, 2020, https://community.esri.com/t5/arcgis-enterprise-portal/where-can-i-find-a-shapefile-with-all-us-counties-and-fips-code/td-p/307592.

[13] "Fast Facts," History, U.S. Census Bureau, accessed November 7, 2020, https://www.census.gov/history/www/through_the_decades/fast_facts/.

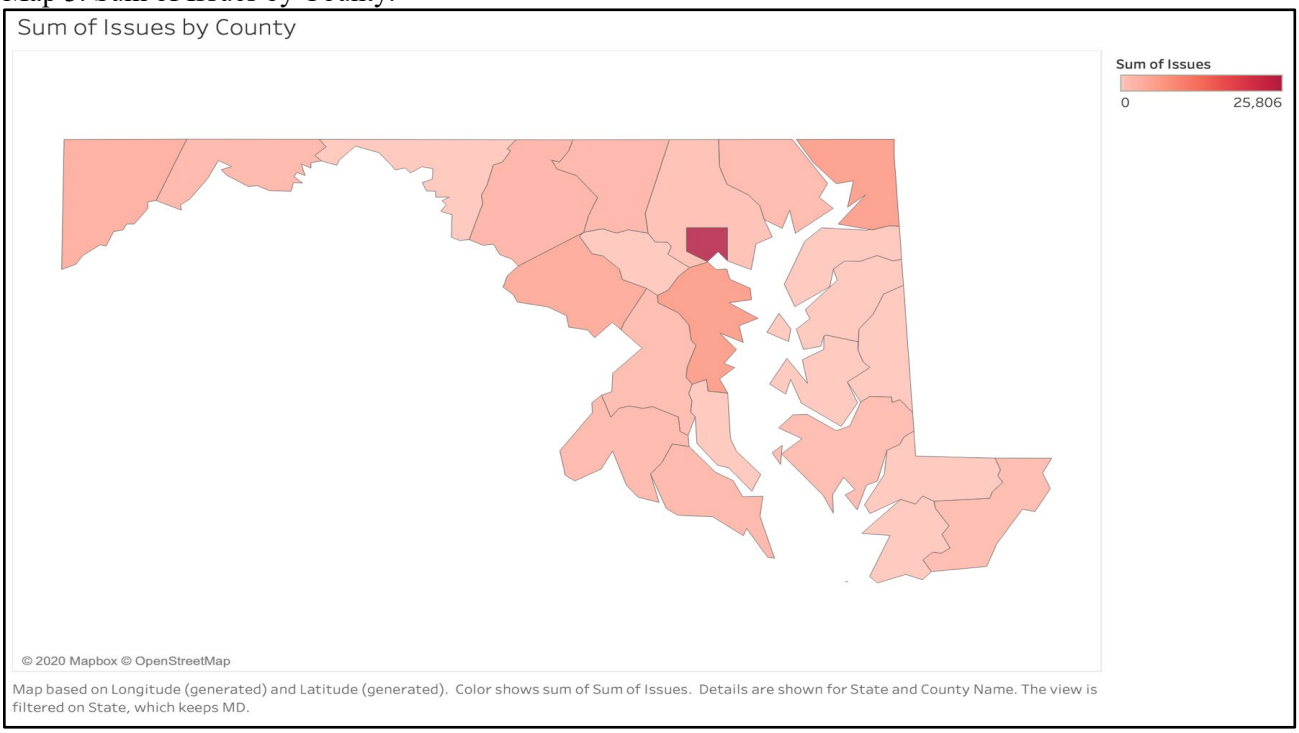
Map 1. Sum of Titles of Underrepresented Groups by County.



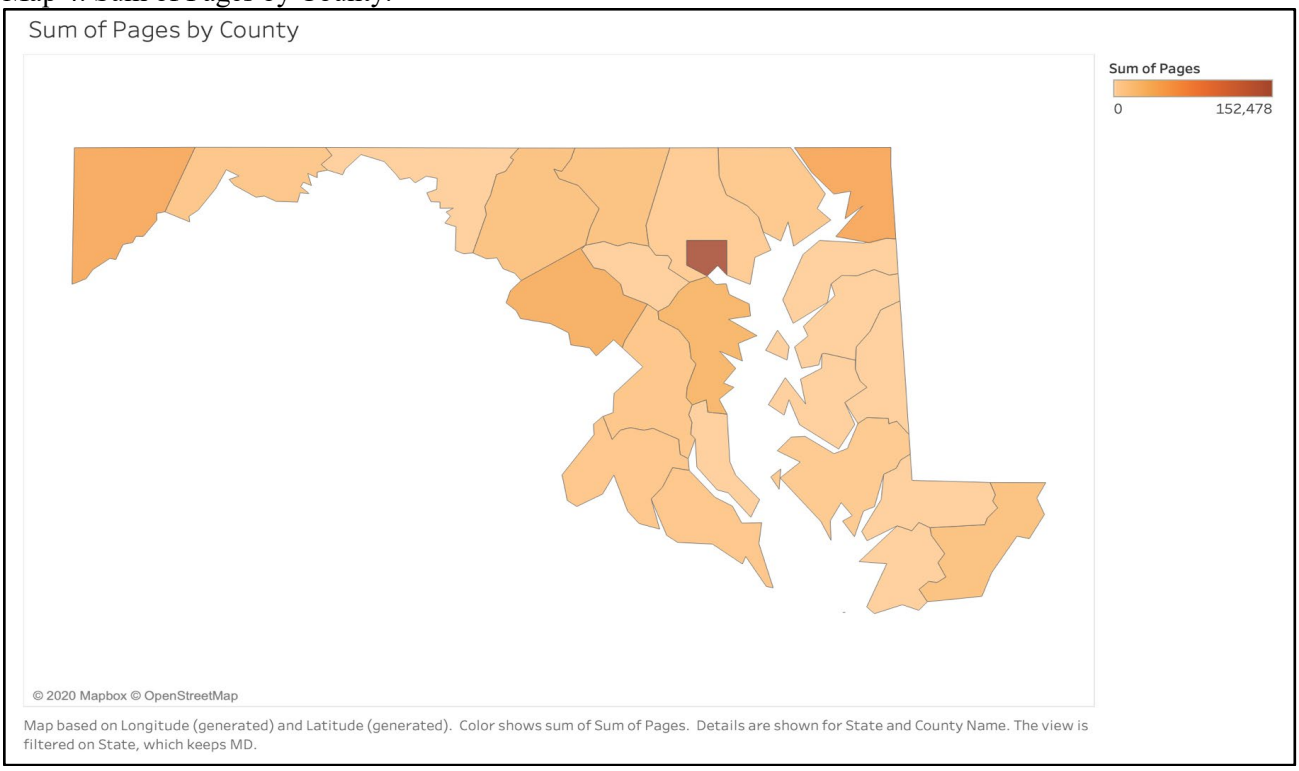
Map 2. Sum of Titles by County.



Map 3. Sum of Issues by County.



Map 4. Sum of Pages by County.



Map 5. Titles, Issues, Pages by County.

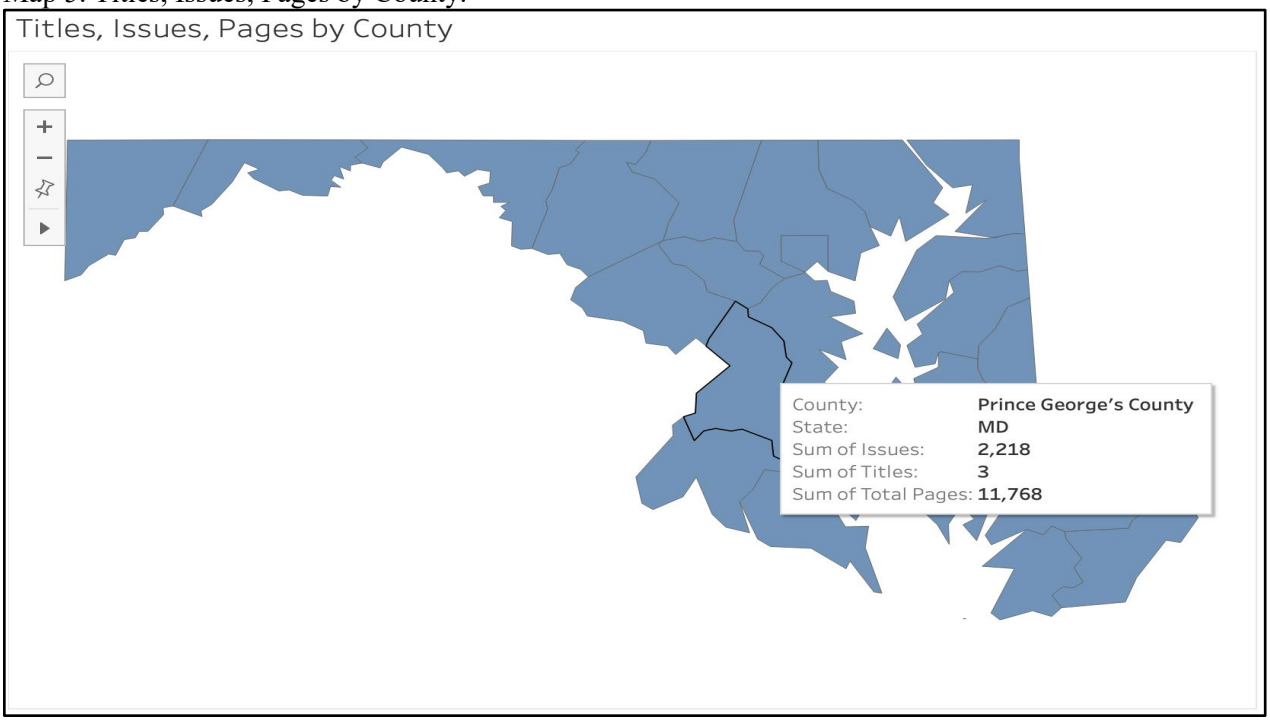
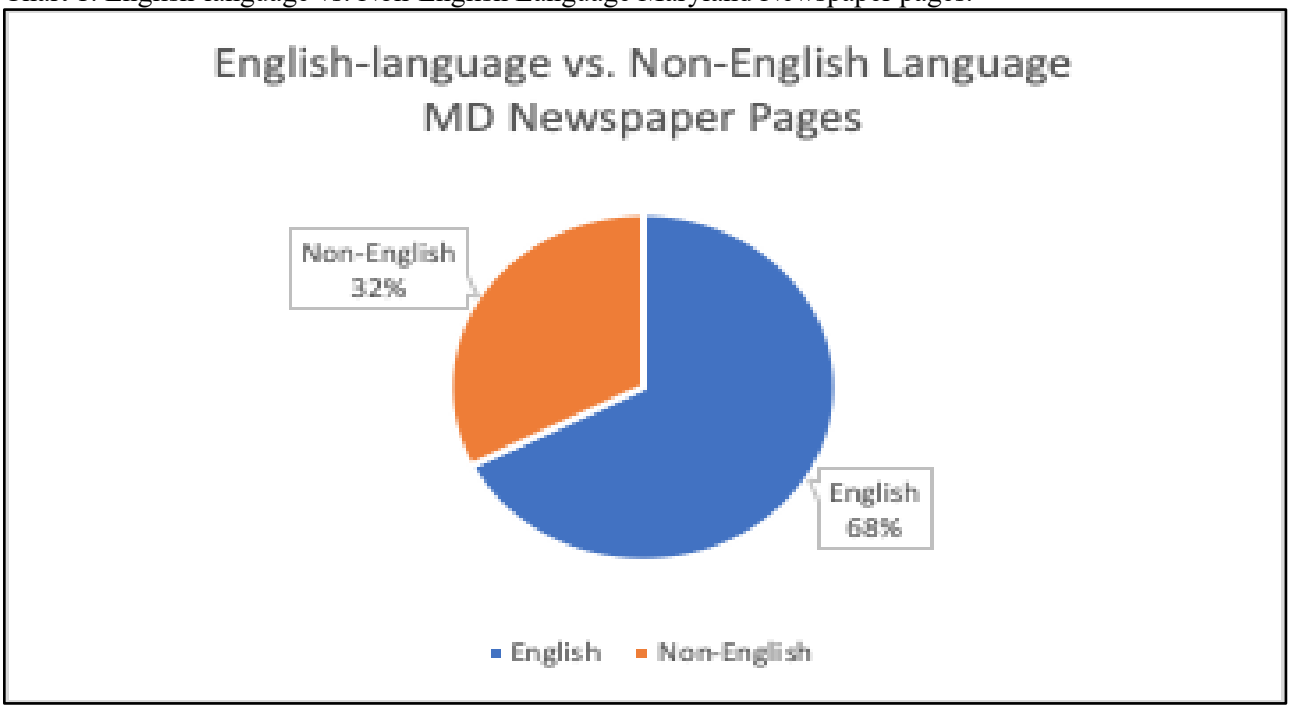


Chart 1. English-language vs. Non-English Language Maryland Newspaper pages.



This data analysis has some limitations and includes some data that is not yet finalized. These considerations should be taken into account when viewing this data:

- HMNP's phase four of NDNP is currently in process. Data from phase four titles that are not yet in Chronicling America are estimates from HMNP metadata collation on what is intended to be submitted to LC, but has not yet been approved by LC.
- Titles, issues, and pages are all important to note because some titles are long runs and others are short, some titles have issues that are only four pages long while others can be upwards of sixteen pages long, and some newspapers are weekly while others are daily. All of this impacts how much content a county or region truly has and why all three sums need to be evaluated.
- Related to title sums, another consideration is the concept of families of titles - newspapers often changed names or merged with other local newspapers and some of these families can be quite large and complicated. This research displays title information as cataloged in the U.S. Newspaper Directory, which includes all official name changes and mergers. This data does not yet show families of titles, which would need to be manually edited in the data spreadsheet. There are fewer overall titles and in some cases a much smaller number of titles for a particular county when families of titles is taken into consideration, but in many instances that would be a more accurate comparison for the purposes of this data analysis and will be a point of further research.
- Since this is historic newspaper content; some cities have changed names and some county borders have changed since the paper was originally published. Modern city names and county borders are used in this data to align with the current geographic data used to create the maps.
- During different phases of the NDNP, LC only accepted newspapers from some certain time periods, though they have now expanded to 1690-1963. With that restriction, it limited which titles could be digitized or at times cut off runs that went longer. This impacts the totals for regions and this is something to consider in the selection decision. A title that was digitized in an earlier phase could be put forth again to finish the run now that the date has been extended.
- Finally, there are a number of reasons why there was no expectation to find equal sums between counties and/or regions of the state. These data visualizations are only meant as a guide for HMNP and researchers and a tool to use in the title selection process. Some reasons for variations include: different newspapers have survived over the years and the quality of the surviving newspaper, or even the instance of a reel(s) of microfilm being unknowingly processed incorrectly previously and losing content from a particular region, papers from a particular county or region having already been digitized by another institution, an NDNP preference toward papers of record which would be from larger cities and limitations of other NDNP selection criteria, migration patterns of immigrant populations, and other unique qualities and characteristics of particular regions of a state that may give it preferential research value.

Findings

While Map 1 doesn't show the whole story of all of the types of underrepresented groups, most of the newspaper titles of underrepresented groups were published in Baltimore. If titles from underrepresented groups outside of Baltimore are identified, these should be considered to receive a broader perspective from diverse communities of underrepresented groups.

In looking at Maps 2 - 5 with sums of title, issues, and pages by county it becomes clear that there are several counties that have no newspaper content at all. As mentioned above, not all counties will necessarily have content for a variety of reasons and certain counties or regions may have preferential research value. However, six counties on the Eastern Shore do not have any titles leaving out representation from an entire region of the Eastern Shore. This should be considered further - have titles from this region been digitized by other institutions? Are there surviving titles from this region? The other counties that do not have content or counties with low counts should be further researched as well. On the

other side, Anne Arundel County has one of the higher counts, but as the state capital, perhaps it should have even further representation, especially since one of the previously digitized titles could be further digitized now after LC extended the accepted date range.

In looking at the non-English language content, it is known that these titles are all from the immigrant communities in Baltimore. Seeing that this is roughly one-third of the total Maryland newspaper content shows that nearly all of the Baltimore content is non-English language. Consideration should be given to more English language content in Baltimore since it is such an important city for the state and country, while balancing that with research of what titles have already been digitized elsewhere that were published in Baltimore.

Conclusion

This analysis is a work in progress that provides some points of consideration for selecting further titles for digitization in phase five including gaps in representation and further research questions. It also begins to paint a picture of the collection for researchers. The interactive map could become a useful feature on the HMNP website for researchers to use as a tool looking for content in the state. In addition to further data analysis, a print and online guide to the collection will be written to describe the Maryland content in Chronicling America and will include some of the analysis information, which will establish a finding aid for this artificial collection.

In October 2020, HMNP held a virtual selection advisory board meeting. Some of this analysis was used in the presentation to the board to share where gaps in the collection may be located and information to consider when choosing the next group of titles for digitization. A preliminary selection list was included with the NDNP application that needs to be narrowed down to 100,000 pages for phase five. This process is ongoing with plans to finalize by December 2020.

Moving forward with analyzing the collection, items to consider include:

- finding other ways of displaying and visualizing the data;
- adding a process to view the data as families of titles;
- determining the most effective way to include temporal data and visualizations;
- continuing research in best practices for diversity and inclusion in digitization selection;
- utilizing data from the digitized newspaper survey that is conducted as a part of the project in a larger analysis; and
- looking for ways to make this process more automated and efficient for the future.

Any additional research will be shared with the HMNP advisory board in making future title selection decisions and the program will continue to evaluate policies and procedures to incorporate diversity and inclusion data in selection and decision making. The HMNP staff and advisory board also will determine the best way to share this information with archives, libraries, and museums throughout the state and make it available on the HMNP website.

This case study could be applied to other NDNP state newspaper projects looking to narrow title selection for their state or other newspaper projects or digital libraries looking for collection analysis and selection criteria to consider for their digitization decisions. This analysis also is especially helpful since this is an artificial collection and other institutions with artificial digital collections may consider this case study useful in understanding these types of collections and making selection decisions for them.

Resources

Dong, Hanyang. "How to build an interactive county level map in tableau." Storybench. School of Journalism. Northeastern University. Accessed November 7, 2020. https://www.storybench.org/build-interactive-county -level-map-tableau/.

Esri. "Uscounties.shp." The Esri Community. Accessed November 7, 2020. https://community.esri.com/t5/arcgis-enterprise-portal/where-can-i-find-a-shapefile-with-all-us-counties-and-fips-code/td-p/307592.

Library of Congress. "Chronicling America." Chronicling America: Historic American Newspapers. Accessed November 7, 2020. https://chroniclingamerica.loc.gov/.

Library of Congress. "Chronicling America Data Visualizations." National Digital Newspaper Program. Accessed November 7, 2020. https://www.loc.gov/ndnp/data-visualizations.html.

Library of Congress. "Content Selection." National Digital Newspaper Program. Accessed November 7, 2020. https://www.loc.gov/ndnp/guidelines/selection.html.

Library of Congress. "U.S. Newspaper Directory, 1690-Present." Chronicling America: Historic American Newspapers. Accessed November 7, 2020. https://chroniclingamerica.loc.gov/search/titles/.

National Endowment for the Humanities. "National Digital Newspaper Program." Division of Preservation and Access. Accessed November 7, 2020. https://www.neh.gov/grants/preservation/national-digital-newspaper-program.

Summerlin, Donnie. "Selecting Newspaper Titles for Digitization at the Digital Library of Georgia." *D-Lib Magazine* 20, nos. 9–10 (September–October 2014). https://doi.org/10.1045/september2014-summerlin.

Tableau. "Maps and Geographic Data Analysis in Tableau." Tableau Help. Accessed November 7, 2020. https://help.tableau.com/current/pro/desktop/en-us/maps.htm.

University of Maryland. "The Historic Maryland Newspapers Project." University Libraries. Accessed November 7, 2020. https://www.lib.umd.edu/digital/newspapers/home.

U.S. Census Bureau. "Fast Facts." History. Accessed November 7, 2020. https://www.census.gov/history/www/ through_the_decades/fast_facts/.

U.S. Department of Agriculture. "County FIPS Codes." Natural Resources Conservation Service. Accessed November 7, 2020. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697.