Submitted on: 01.06.2017

2017 IFLA International News Media Conference
27-28 April
Reykjavík, Iceland

Rediscovering Music Printed in American Newspapers: A Case Study for Mining Non-Textual Content from Digital Newspapers

Sarah Lynn Fisher
Digital Newspaper Unit, University of North Texas Libraries, Denton, Texas, U.S.A.
E-mail address: sarahlynn.fisher@unt.edu

Abstract:

From approximately 1900 to 1920, music publishers like the American Melody Company provided a syndicated "song service" to newspapers in diverse regions across the United States. Solo piano pieces, solo songs with piano accompaniment, and orchestral reductions of instrumental pieces were rotated among the subscribing newspapers and published with little commentary other than the publishing copyright. This song service highlights how newspapers served as a disseminator of art and culture prior to the widespread availability of other media formats. Music historians might use the music in these newspapers to explore early 20th century American popular and parlor music traditions.

Many of these compositions are no longer extant beyond their publication in newspapers, which are now preserved on microfilm. Digital newspaper databases, such as the Library of Congress' open access Chronicling America: Historic American Newspapers*, have the potential to rescue this music from obscurity by hosting them online in a digital format. However, searching within* Chronicling America *requires a text-based search term for user inquiries, as searches are conducted based on the text file associated with each newspaper image that is derived via Optical Character Recognition (OCR) software. When it is only possible to search the text on the pages, not images or musical notation, researchers must employ alternative strategies to locate non-textual content. Using music printed in newspapers available on* Chronicling America *as a case study, these methods include metadata conversion into a machine-readable format (JSON) using the Chronicling America application programming interface (API) and analyzing publishing trends using visualizations created using this metadata.*

Keywords: Digital Newspapers, Music Indexing, OCR Analysis, Data Visualization, *Chronicling America*

Introduction

Massive archives of digital content present both opportunities and challenges for research in the digital humanities. These archives allow users to consider an entire corpus of historical documents, instead of a representative subset, and "expand the kinds of research questions that can be pursued" (Audenaert & Houston, 2013, p. 9). Text remains the primary search and retrieval method of digital libraries, whether through the application of optical character recognition (OCR) software or textual metadata associated with a digital image. Text-based queries limit the types of searches that can be performed within cultural datasets thereby restricting their research potential.

News media archives include a variety of multimedia content that represents the social, cultural, and political histories of specific populations. Newspapers in particular served as a disseminator of art and culture prior to the widespread availability of other media formats. For example, music publishing companies provided sheet music to newspaper publishers to include in editions of their newspapers. Subscribers with pianos in their homes could collect these into a songbook and play popular songs of the time, allowing them to hear live music at home even if they did not own a phonograph. Music historians today might use the music in these newspapers to explore early 20th century American popular and parlor music traditions. This research would be especially valuable as almost all of these compositions and the names of their composers have disappeared from cultural memory since their publication in newspapers.

Digital newspaper databases have the potential to rescue this music from obscurity and allow it to enter the scholarly discourse on early twentieth century popular American music. *Chronicling America: Historic American Newspapers*, the Library of Congress' (LC) open-access digital newspaper database, is an open-access digital repository of newspapers digitized from microfilm that is funded by the National Endowment for Humanities (NEH) and maintained and hosted by the LC. The NEH awards grants to institutions in each state to support digitization and metadata creation activities as part of the National Digital Newspaper Program (NDNP); awardee institutions in turn submit a specified number of digitized newspaper pages to LC for inclusion in *Chronicling America* during the two-year grant award period. The submitted content must conform to technical guidelines stated by LC and should also fall within a specific date range. For digitization grants awarded prior to 2016, LC only accepted newspapers that are currently in the public domain. *Chronicling America* currently includes nearly 11.8 million digital newspaper pages from 44 states and territories.

Content such as notated music in newspapers is of interest to researchers but is not represented in OCR files associated with the digital image and therefore is not directly accessible via the search methods offered by most digital repositories. While these repositories have increased the availability of this content, they have not increased its accessibility for users with non-text based research problems. Research teams such as Audenaert and Houston (2013); Herbert, D., Palfray, T., Nicolas, S., Pranouez, P., and Paquet, T. (2014); and Lorang, Soh, Datla, and Kulwicki (2015) have been exploring this issue using image segmentation and machine learning, but how can a user with more limited technical skills locate and access non-text content in a digital repository? Using music printed in American newspapers available in *Chronicling America* as a case study, the research value of developing tools to locate non-textual content for users of digital repositories is immediately apparent.

Literature Review

Recent studies have identified both the potential of content-based retrieval for newspapers and possible avenues for using newspapers in musicological and digital humanities research.

Lorang et al. (2015) identify "a critical gulf between the amount of material that is available to researchers and the ability researchers have to find the materials they need within digitized collections of primary materials" (para. 4). Multimedia content, such as images or musical notation, represents "unstructured" information (Castelli, 2009, p. 5024). This means that the information messages this media conveys cannot be easily quantified and displayed in written language. Castelli (2009) argues that while digital imaging methods have advanced, and replaced traditional imaging in many applications, image retrieval has not "kept up with the pace of the digital imagery explosion," relying instead on keyword-based searches and image indexes organized by descriptive metadata (p. 5022). The disconnection between retrieved search results and the user's search objectives has been identified as a "semantic gap" in information retrieval (Castelli, 2009, p. 5025). This observation is echoed in Lorang et al. (2015), who contend that "one way in which the basic functionality of digital libraries has stalled is that text nearly always serves as the primary, and most often the only, basis for retrieval and analysis in conventional systems" (para. 5).

In some research situations, text-based queries are an inadequate method of searching what is primarily a text-based medium, such as books and newspapers. In the case of both Audenaert & Houston (2013) and Lorang et al. (2015), both teams were concerned with identifying and analyzing the visual attributes of poetry in large cultural datasets. The discovery systems developed by their respective research projects, VisualPage and Image Analysis for Archival Discovery (Aida), are indebted to the methods of image segmentation and analysis described in Herbert et al. (2014). Herbert et al. (2014) developed a method to identify the elements and layout of a newspaper page, including articles, titles, sub-titles, paragraphs, figures and captions (p. 5). This labeling begins at the pixel level in which the image is first segmented with the help of a Conditional Random Fields (CRF) modeling algorithm. Each pixel is associated with a logical label, such as background, title character, vertical separator, and noise. These entities are then associated with the higher-level organization of the newspaper page (the attributes described above) through comparison with statistical models (Herbert et al., 2014, p. 4-5). This type of "smart indexing and retrieval" is necessary due to both the size of digital newspaper databases (which can contain millions of pages) and the frequently degraded condition of the documents (Herbert et al., 2014, p. 3).

Lorang et al. (2015) use similar methods of supervised machine learning to locate poetry published in newspapers that have been digitized for *Chronicling America* (para. 2). A training set of 210 images, each snippets of a newspaper page, is processed using a series of algorithms to isolate the visual structure of poetry within the image. These blurring, binarization, and consolidation algorithms reduce the snippets to their basic "visual cues" (Lorang et al., 2015, para. 14). This dataset is used to test a collection of images for the presence of poetic content, resulting in a result of either "true" or "false" (Lorang et al., 2015, para. 24). While the results necessitate some refinement of the test attributes, the Aida system represents an important step towards the integration of content-based retrieval (CBR) methods within digital newspaper databases.

Using the same literary genre as a basis for their prototype, Audenaert and Houston (2013) developed the VisualPage system, which extracts the "visual and bibliographic aspects of printed texts" to aid in analysis of their cultural meaning (p. 9). Just as the nineteenth-century poetry books in the initial dataset convey meaning on three levels (bibliographic, visual, and linguistic), VisualPage analyzes cultural heritage materials on three levels: the visual structure of individual page, pattern analysis across the dataset, and visualizations of data through an interactive interface (Audenaert & Houston, 2013, p. 9, 11). The goal of this type of research

is to “understand the history of visual design; spot trends and influences; and pose questions over the entire corpus of nineteenth-century poetry, rather than with only a few canonical texts” (Audenaert & Houston, 2013, p. 12).

Digital newspaper archives of the quality and scope of *Chronicling America* have only appeared in the last decade, but musicologists and other scholars are beginning to recognize the ways in which digital newspaper collections can enable research and expose hidden and forgotten areas of history. Nicholson (2013) summarizes this potential and demonstrates how new digital research methodologies might be applied to digital newspaper archives. Through full-text keyword searching, Nicholson concludes, “billions of individual words—the fundamental building blocks of culture—are now at our fingertips” (p. 67). Crist and Marvin (2008) mention sheet music published in newspapers as an emerging area of archival research in their introduction to the recent collection *Historical Musicology: Sources, Methods, and Interpretation*. Ferris (2011) has expanded on this kernel of an idea in a recent dissertation from the National University of Ireland, which explores musical life in Dublin in the 1840s through the lens of published newspapers. In the realm of indexing content in newspapers, Michael Tilmouth’s *A Calendar of References to Music in Newspapers Published in London and the Provinces (1660-1719)*, originally published in the *Royal Musical Association Research Chronicle*, is a notable example. Tilmouth amassed a vast quantity of information from newspapers during his research and saw the benefit of publishing it to future scholarship. A similar index of music published in American newspapers would allow researchers pursue in-depth research questions related to the music and its social context because the entire corpus of this content would be available to them.

Rediscovering Music Printed in Historic American Newspapers

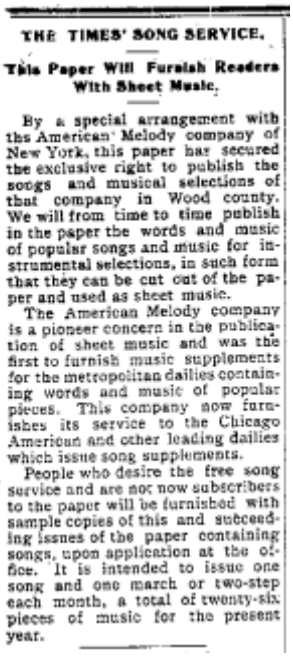
Beyond the sources mentioned above, little to no research has been done into the extent and scope of music printed in American newspapers in the early 20th-century. This perhaps might be a result of the fact that microfilming has been the dominant method of newspaper preservation since this technology was introduced to libraries (Baker, 2001). A user who encountered music in a newspaper while reviewing the microfilm might be intrigued enough to look for more examples but would find it time-consuming to scroll through reels and reels of microfilm to locate them, especially when no definitive list of dates and newspapers titles that published sheet music exists.

Digital newspaper repositories like *Chronicling America* have made searching and browsing much easier but only when a known text string can be used to initiate the search. Without the availability of tools such as VisualPage or Aida systems described above, a user must approach researching non-text-based content in newspapers from other angles. Searches employing common musical terms, such as "*adagio*" or "*D.C. al fine*," yield either too many unrelated results (10,024 for "*adagio*") or more precise results that do not include all possible examples (69 for "*D.C. al fine"*). Analyzing the results of the latter search, it is apparent that a publisher's copyright statement often accompanies the printed music. Two of the music publishers represented among the results in *Chronicling America* are the Murray Music Company (776 results) and the American Melody Company (398 results). (Shortening the word "company" to "co" yields more results because this word is sometimes abbreviated in the copyright statement.) Categorizing these results one by one is possible, but other more programmatic methods might illuminate trends much quicker. Using one of these publishers as a case study, the American Melody Company, and the application programming interface (API) available

on the *Chronicling America* website certain aspects of this music's social and cultural context come into view.

The American Melody Company: A Case Study

On April 21, 1905, a short article (Figure 1) appeared on page twelve of *The Marshfield Times* (the self-proclaimed "Largest Paper in Central Wisconsin") entitled "The Times' Song Service:"



THE TIMES' SONG SERVICE.

This Paper Will Furnish Readers With Sheet Music.

By a special arrangement with the American Melody company of New York, this paper has secured the exclusive right to publish the songs and musical selections of that company in Wood county. We will from time to time publish in the paper the words and music of popular songs and music for instrumental selections, in such form that they can be cut out of the paper and used as sheet music.

The American Melody company is a pioneer concern in the publication of sheet music and was the first to furnish music supplements for the metropolitan dailies containing words and music of popular pieces. This company now furnishes its service to the Chicago American and other leading dailies which issue song supplements.

People who desire the free song service and are not now subscribers to the paper will be furnished with sample copies of this and succeeding issues of the paper containing songs, upon application at the office. It is intended to issue one song and one march or two-step each month, a total of twenty-six pieces of music for the present year.

Figure 1: Newspaper clipping titled "The Times' Song Service"

From approximately 1900 to 1920, the American Melody Company provided this syndicated "song service" to newspapers in spread across the United States, including *The Washington Bee* (Washington, D. C.), *The Bourbon News* (Paris, Ky.), *The Chickashaw Daily Express* (Chickashaw, Okla.), *The Labor World* (Duluth, Min.), and the *Scranton Wochenblatt* (Scranton, Pa.). Solo piano pieces, solo songs with piano accompaniment, and piano reductions of instrumental pieces were rotated among the subscribing newspapers and published with little commentary other than the universal annotation of "Copyright by the American Melody Company, New York."

A survey of this published music available in the digital newspaper archives *Chronicling America* suggests a few commonalities among the repertoire; solo piano music is often either short character pieces with descriptive titles, such as John A. Allen's "Dance of the Banshees"[1] and "The Whirl-Wind Gallop" by Jacques Mendelsohn,[2] or piano reductions of instrumental works from popular and art music genres. Examples of these include "'The Bathing Girls' Two

[1] *The labor world.* (Duluth, Minn.), 12 October 1907, 5. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn78000395/1907-10-12/ed-1/seq-5/.

[2] *The Chickasha daily express.* (Chickasha, Indian Territory [Okla.]), 01 March 1912, 6. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn86090528/1912-03-01/ed-1/seq-6/.

Step As Danced by the Hollis Sisters in vaudeville"[3] and "'Lover's Vision' (*Libestraum*) Reverie. As played by the Plaza Hotel Orchestra, New York."[4] (For the latter two, no composer information is given.) Songs usually have descriptive, often sentimental, titles and may include piano accompaniment: "Among the Flowers" by H. G. Allaire and Harry A. Stewart[5] and "'Some One' Song. Words by Lulu Irene Brown. Music by Frederic Preston." are two examples.[6] Notable for its continued recognition in posterity is the publication of "'Dance of the Hours' from Ponchielli's Opera 'La Gioconda' Sung by Enrico Caruso at the Metropolitan Opera House N. Y."[7]

As was noted in the two instrumental examples cited above, the published music did not always credit the composer and none of the works surveyed credit an arranger. Always included is a copyright attribution to the "American Melody Company, New York" (sometimes with the abbreviations "Co." and "NY"). Titles did, however, often include references to a performer or performing group and the location of a presumably well-known performance of a work. Thus, newspapers subscribers with pianos in their homes would be able to learn and reproduce versions of popular or famous compositions that they might have only read about in the newspaper or heard in local concerts. An opportunity to travel to New York from Baxter Springs, Kansas to New York City to hear the Plaza Hotel Orchestra would be rare, if not impossible, in 1910.

The syndicated sheet music service of the American Melody Company continued until approximately 1920.[8] While the circulation statistics of subscribing newspapers varied widely, one might assume that millions of Americans received copies of the compositions cited above, particularly when they appeared in more than one newspaper. Subscribers might have clipped out this published music from the newspaper (as was suggested in *The Marshfield Times* article quoted above) and kept it as part of their sheet music collections while discarding the other more ephemeral pages and sections. What is curious, then, is how almost all of these compositions and composers have disappeared from memory since their publication in newspapers; none of the examples cited above, excepting the Ponchielli aria, have been published elsewhere, and their composers are not included biographical dictionaries. In addition, song titles are absent from comprehensive American popular song indexes. An index or anthology of music published in newspapers by the American Melody Company has not been published.

---

[3] *The Washington bee.* (Washington, D.C.), 18 March 1911, 2. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn84025891/1911-03-18/ed-1/seq-2/.

[4] *Baxter Springs news.* (Baxter Springs, Kan.), 10 March 1910, 4. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn83040592/1910-03-10/ed-1/seq-4/.

[5] *The News-Herald.* (Hillsboro, Highland Co., Ohio), 27 August 1908, 3. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn85038161/1908-08-27/ed-1/seq-3/. Also appeared in: *The Bourbon news.* (Paris, Ky.), 01 January 1909, 3. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn86069873/1909-01-01/ed-1/seq-3/.

[6] *The Butler weekly times.* (Butler, Mo.), 09 October 1902, 2. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn89066489/1902-10-09/ed-1/seq-2/.

[7] *Der tägliche Demokrat.* (Davenport, Iowa), 21 January 1917, 3. *Chronicling America: Historic American Newspapers*. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn84027107/1917-01-21/ed-1/seq-3/.

[8] The earliest example surveyed in *Chronicling America* is 1902, and the latest is 1917. A brief mention in the *New York Times* indicates that the American Melody Company was incorporated in New York State in 1902 with an initial capital of $20,000 and under the direction of J. A. Shay, I. R. Smith, and Ferdinand Wetzler. "New York Incorporations," *New York Times* (New York), 01 June 1902, 18. ProQuest LLC, *ProQuest Historical Newspapers: The New York Times (1857-1922) with Index.*

Analyzing Publishing Trends

As stated above, there are currently 398 results for the search term "American Melody Co" in *Chronicling America*. Keeping in mind the limitations of the repository itself[9], visualizations created from metadata associated with the newspaper pages included in the search results will help researchers to quickly answer questions about this music printed in newspapers, including: what newspaper titles printed music provided by the American Melody Company and where were these newspapers published? What page of the newspaper did the music most often appear on? During what years was the syndicated music from the American Melody Company published in newspapers and what were the most popular years in the date range? Answering these questions using visualizations will bring focus to future research efforts outside of *Chronicling America* for music printed in newspapers, saving research time during what might otherwise be a tedious search process wading through some of the peak years in American newspaper publishing.

The visualizations for the American Melody Company search results shown below were produced using the *Chronicling America,* Microsoft Excel spreadsheet software, and the data visualization software, Tableau. The publicly-available API behind the *Chronicling America* website utilizes the OpenSearch protocol. Thus, any search performed within the *Chronicling America* graphical user interface (GUI) can be viewed in a different format by modifying the URL query string. The search query for "American Melody Co" in the GUI returns a URL in HTML format by default, but can be viewed in JSON format by adding "&format=json" to the end of the URL query string. JSON format is useful for creating visualizations because the metadata elements are separated by commas; thus, using a parsing script, the metadata can be converted to a table format that is compatible with a data visualization software. Viewing the search results for "American Melody Co" in JSON format, it is easy to see that there are 398 total results, but only 20 results are displayed per page. To view all of the results on a single page, the number of rows in the URL string query can be changed to a larger number. Changing "&rows=20" to "&rows=400" in the URL displays all of the possible results on a single page. Using a method described by Gambill (2014), which is based on code written by Lohrbeer (2014), the search results in JSON format can be converted to a Google spreadsheet. This method utilizes the Script Editor tool available in Google Sheets. The spreadsheet can then be downloaded in the Microsoft Excel spreadsheet format, .xlsx, which is easily imported into a data visualization software such as Tableau.

Cleaning the Data

Once the metadata is converted to a table format in Excel, the metadata fields that will be used to create the visualizations should be examined for any outliers. This can be done using the Filter tool in Excel. In this dataset, the "Items City," "Items Date," "Items Title," "Items State," and "Items Place of Publication" fields were reviewed using the Filter tool. For the locational and title metadata fields, any unusually formatted strings that might confuse the data visualization software were changed. For example, "Mississippi,Mississippi,Mississippi" should be corrected to "Mississippi" in the "Items State" field. In addition, values that might be distracting in the visualization should also be altered. Several title values included the word "volume" following the newspaper title in the "Items Title" field. These values were changed

[9] *Chronicling America* includes only specific newspaper titles chosen by institutions in states that have been awarded digitization grants. Additionally, due to the quality of the microfilm from which the digital newspaper images are derived, the OCR file is often inaccurate.

to exclude the word "volume" from the title; for example "Scranton Wochenblatt. volume" was corrected to "Scranton Wochenblatt." Once a group of values is identified, the Find and Replace tools in Excel can be used to quickly change all of the same values to the standardized version.

The "Items Date" field proved particularly challenging to reformat. In the *Chronicling America* dataset, this field was formatted as a string of numbers in the "YYYYMMDD" format. This was interpreted by Excel as a string and not a date format. However, formatting the cells within that particular column as a Date category within Excel was not sufficient to allow the data visualization software to recognize the values in this field as a Date data type. Several other options were explored, including the DateParse function in Tableau and the data cleaning software OpenRefine. Finally, all that could be done was to add dashes between the year, month and day numbers within Excel. With the date in the "YYYY-MM-DD" format, the software was able to recognize the values in this field as a Date data type. Additionally, the Filter tool in Excel was used to find values in the dataset prior to 1900. Only two such values were included in the set; upon further review of the corresponding digital objects in *Chronicling America*, these newspaper pages were not found to include any sheet music, and thus were added to the set due to OCR errors. These two rows were then removed from the spreadsheet in Excel.

Visualizations and Analysis

The visualizations were created using Tableau Professional Edition (Version 10.0) visualization software. Tableau was chosen because it allows for easy customization of visualizations from default settings that are clean and visually compelling. Additionally, the automatically generated measure "SUM(Number of Records)" in Tableau is useful for the visualizations presented here, as this corresponds to the number of newspaper issues. Adding the number of issues to the visualizations allows the charts and maps to show trends among the nearly 400 digital objects represented by the dataset.

The first visualization (Figure 2) is a filled line chart. To create this visualization, the "Items Date" field was placed on the "Columns" shelf and the "SUM(Number of Records)" was placed on the "Rows" shelf. With this chart and all of those following, the axis labels and titles for each chart or map were changed to provide a clearer representation of the values that are being measured. For example, "Number of Records" was changed to "Number of Issues" and "Items Date Year" was changed to "Year of Publication." Additionally, every effort was made to conform to a uniform font and color scheme. Figure 2 clearly shows that the syndicated music publication service provided by the American Melody Company to newspaper publishers clearly peaked between the years 1907 and 1912. If a researcher is looking outside of *Chronicling America* for music published in newspapers by the American Melody Company, one might begin by looking a newspapers published in this data range.

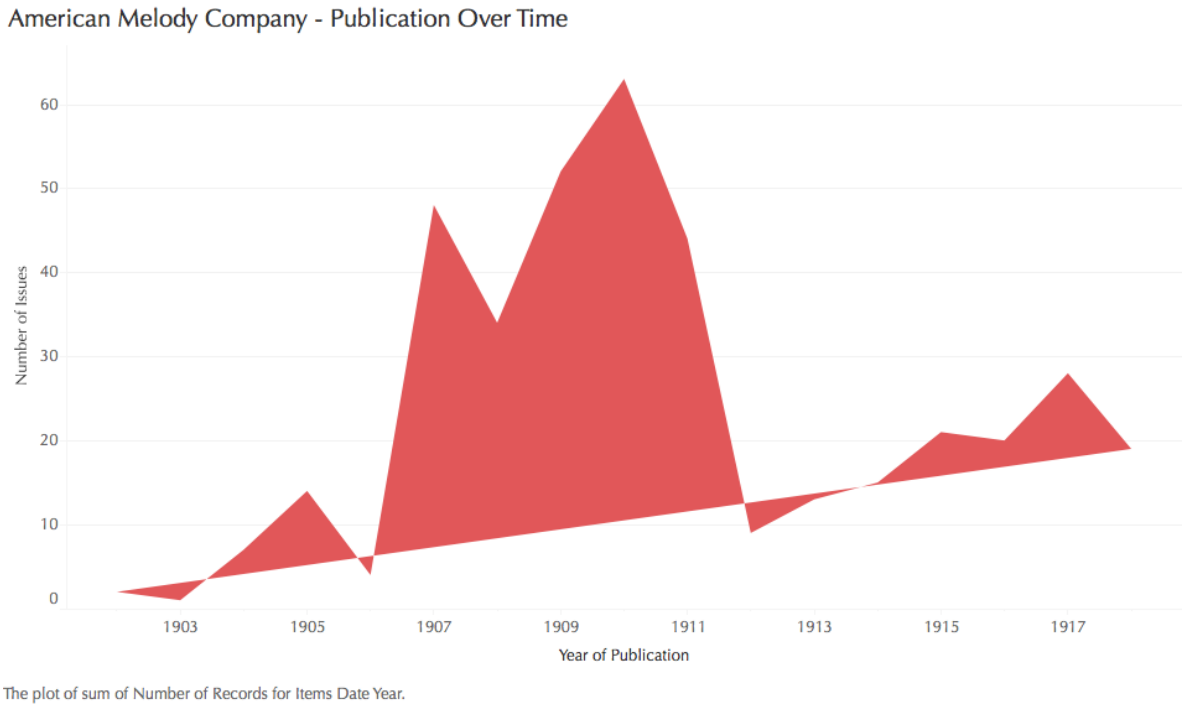


Figure 2: Filled Line Chart of Year of Publication (x-axis) and Number of Issues (y-axis)

Figure 3 is a filled bar graph created by adding the "Page In Sequence" field to the Columns shelf and the "SUM(Number of Records)" measure on the Rows shelf. This visualization clearly shows that the third, second, and sixth pages were the most common pages for the sheet music to appear within the newspaper issue. This might save researchers time when reviewing newspapers in other digital repositories or on microfilm. Most often, one would not need to look beyond the third page of the issue to find out if sheet music from the American Melody Company is included in that particular issue.

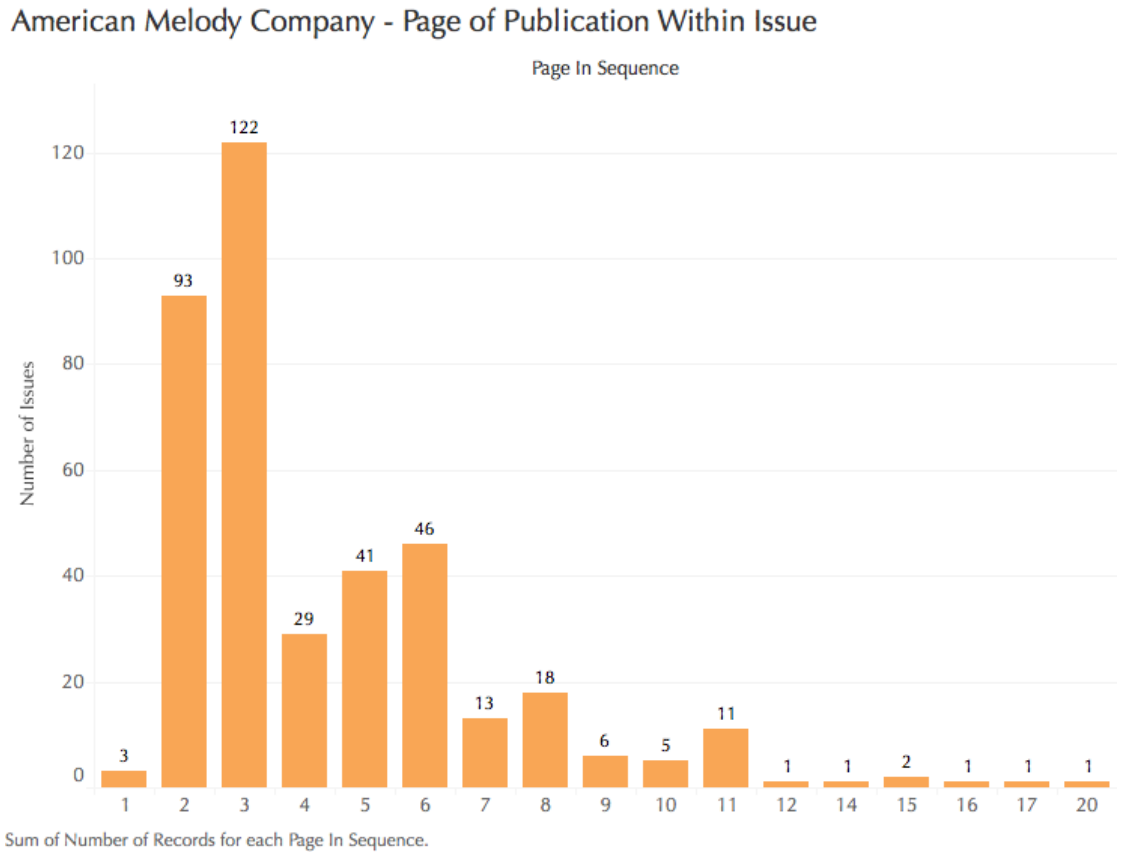
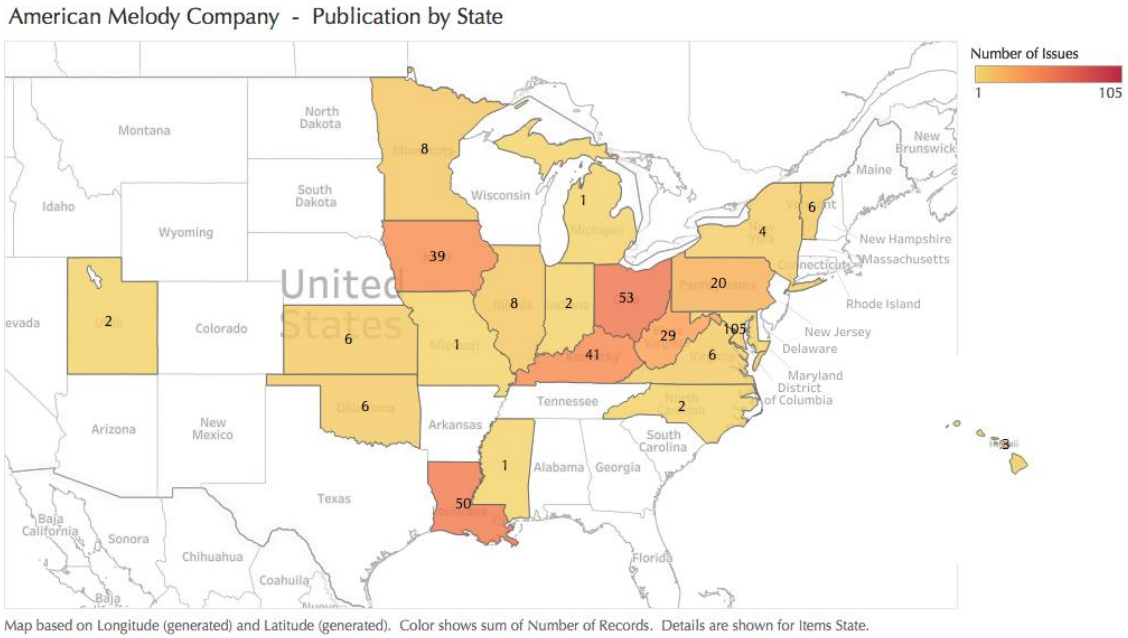


Figure 3: Filled Bar Graph of Page in Sequence (x-axis) and Number of Issues (y-axis)

The final three visualizations examine the newspapers titles included in *Chronicling America* that published sheet music provided by the American Melody Company and the place of publication for these titles. Figure 4 is a map showing the states where newspapers with American Melody Company music were most often printed. This filled map was creating by adding the "Items State" field to the Label mark and the "SUM(Number of Records)" field to the Color mark. States not shown in the western half of the continental United States and Alaska did not have any newspapers included in the dataset. Additionally, in order to present this map in sufficient detail in this paper, the state of Hawaii was superimposed onto a cropped version of the map using Adobe Photoshop.

From this map, it is easy to see that the syndicated song service was most popular among newspaper publishers in the Eastern and Central United States, with Washington, D. C. having the most issues including sheet music, followed by Ohio, Louisiana, and Kentucky. Other newspapers published during a similar time period from these locations might be examined to determine if these titles also include sheet music published in the issues.



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Records. Details are shown for Items State.

Figure 4: Map of United States showing Number of Issues Published in each State

Figures 5 and 6 examine the results from the map in further detail using a heat map and a highlighted table, respectively. These type of visualizations can include more text, such as the newspaper title, without appearing cluttered. Figure 5 was created by adding both the "Items Title" and "Place of Publication" fields to the Rows shelf. The visualization is further enhanced by adding "SUM(Number of Records)" to the Label and Color marks. Within the heat map, the results are sorted so that the title with the greatest number of issues included in the dataset appears at the top of the chart. This chart shows researchers which titles represented in the map in Figure 4 contributed the greatest number of results to the dataset. As expected these, titles were published in the states of Washington, D. C., Ohio, and Louisiana. Researchers might focus on the titles with the greatest results to look for sheet music published by the American Melody Company.

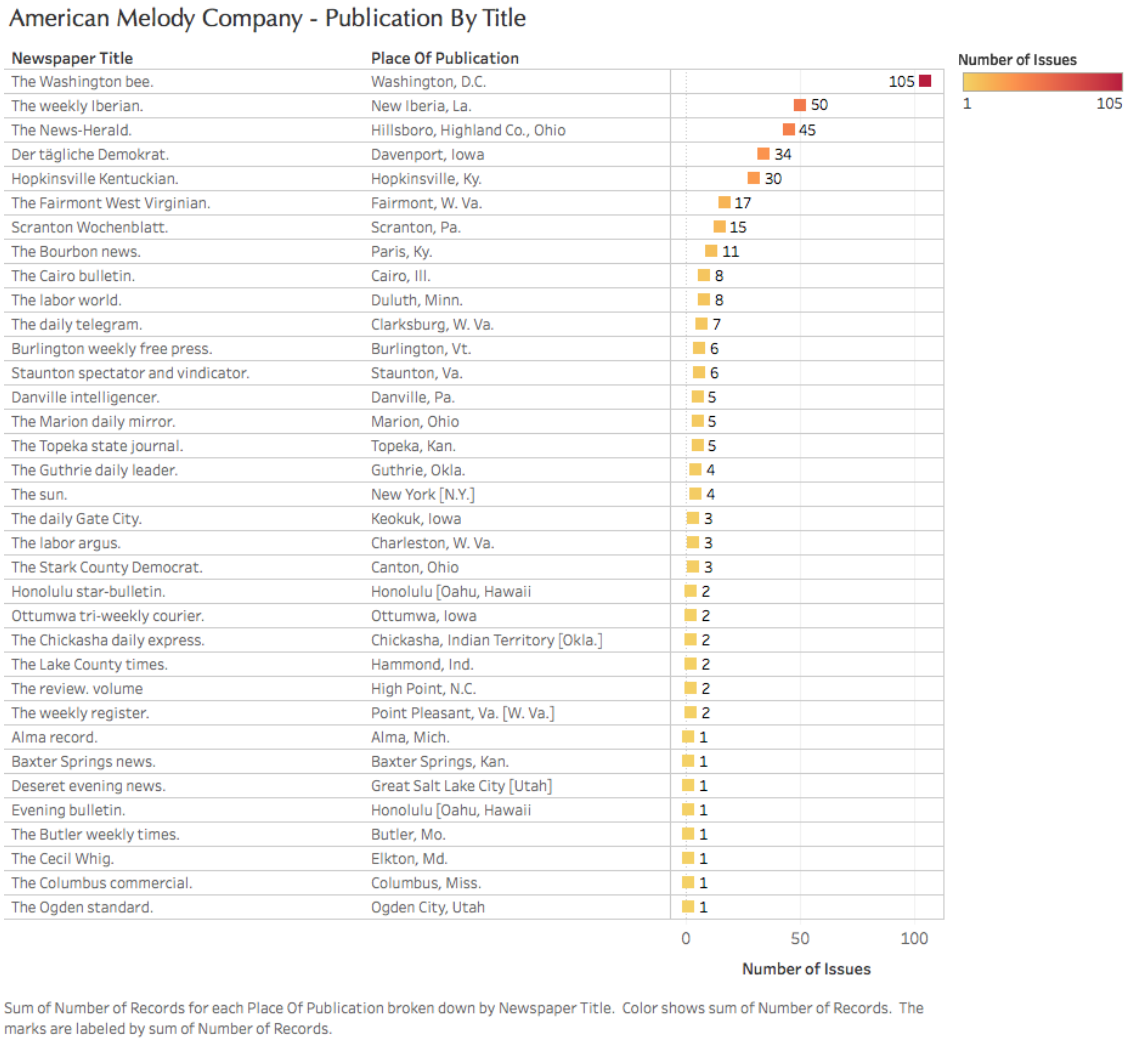


Figure 5: Chart showing Newspaper Title, Place of Publication, and Number of Issues

Figure 6 shows the year that these newspaper titles with the greatest number of results were most often published. This filled table was created by adding the "Items Title" field to the Rows shelf, the "Items Date" field to the Columns shelf, and the "SUM(Number of Records)" to the Label and Color marks. As predicted from the results in Figure 1, the title with the greatest number of results, *The Washington Bee* (Washington, D. C.), included sheet music published by the American Melody Company between the years of 1907 and 1912. This chart does not present as clear a picture of the trends in the dataset. Titles that contributed only one result to the dataset might be further examined to determine if they were included due to OCR abnormalities. However, researchers looking for sheet music examples might focus their efforts on the top seven titles in on the chart for the highest probability of finding accurate results.

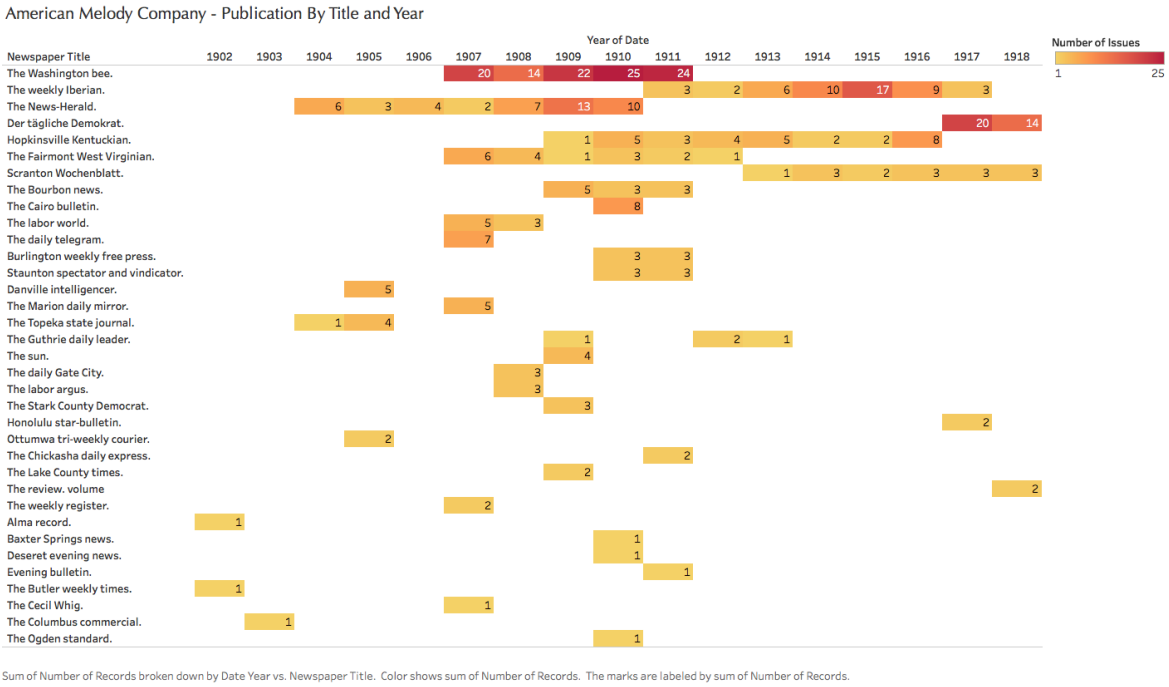


Figure 6: Chart showing Newspaper Title with Number of Issues Published by Date

Conclusion

The visualizations shown above represent only the beginning of what can be done to programmatically index the contents of this rich dataset. Future studies could to programmatically compare the metadata in JSON format to the digital newspapers themselves. However, it should be noted that the analysis done here is still the result of a text-based search for words that happened to accompany musical notation printed in the newspapers. There could very well be other instances of musical notation in newspapers included in *Chronicling America* that are not included in the text-based searches mentioned previously. Other non-textual content in newspapers may not have any captions or text associated with that could be located using the searching of OCR text that is currently available for users of *Chronicling America*. Locating this content will require more robust content-based retrieval systems such as VisualPage or Aida. If digital cultural heritage repositories could find ways to incorporate non-text based searches, they would make discoverable yet another layer of rich content that already exists in these collections for their users.

References

Audenaert, N. & Houston, N. M. (2013). VisualPage: Towards large scale analysis of nineteenth-century print culture. *2013 IEEE International Conference on Big Data*, 9-16. doi:10.1109/BigData.2013.6691665

Baker, Nicholson. (2001). *Double fold: Libraries and the assault on paper*. New York: Random House.

Castelli, V. (2009, Dec 9). Still image search and retrieval. In M. Bates & M. Maack (Eds.)

Encyclopedia of Library and Information Sciences (Third ed., pp. 5022-5041). Taylor and Francis: New York. doi:10.1081/E-ELIS3-120044400

Crist, S. A. & Marvin, R. A. (Eds.). (2008). *Historical musicology: Sources, methods, and interpretation*. Eastman Studies in Music. Rochester, NY: University of Rochester Press.

Ferris, C. (2011). *The use of newspapers as a source for musicological research: A case study of Dublin musical life 1840–44.* (Doctoral dissertation). Retrieved from http://eprints.maynoothuniversity.ie/2577/

Gambill, P. (2014, March 3). How to import JSON data into Google spreadsheets. Retrieved from https://medium.com/@paulgambill/how-to-import-json-data-into-google-spreadsheets-in-less-than-5-minutes-a3fede1a014a

Herbert, D., Palfray, T., Nicolas, S., Pranouez, P., & Paquet, T. (2014). Automatic article extraction in old newspapers digitized collections. *DATeCH '14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 3-8. doi:10.1145/2595188.2595195

Library of Congress. (2017). The National Digital Newspaper Program (NDNP) technical guidelines for applicants [PDF document]. Retrieved from http://www.loc.gov/ndnp/guidelines/NDNP_201719TechNotes.pdf

Library of Congress. (2017). Chronicling America: About the site and API. Retrieved from http://chroniclingamerica.loc.gov/about/api/

Lohrbeer, T. (2014, September 29). google-docs [Github repository.] Retrieved from https://github.com/fastfedora/google-docs/blob/master/scripts/ImportJSON/Code.gs

Lorang, E., Soh, L.-K., Datla, M. V. & Kulwicki, S. (2015, July/August). Developing an image-based classifier for detecting poetic content in historic newspaper collections. *D-Lib Magazine, 21*(7/8). doi:10.1045/july2015-lorang

Nicholson, B. (2013). The digital turn: Exploring the methodological possibilities of digital newspaper archives. *Media History*, *19*(1): 59-73.

Tilmouth, M. (1962). A calendar of references to music in newspapers published in London and the Provinces, 1660-1719. *Royal Music Association Research Chronicle, 2*(1): 1-15. doi:10.1080/14723808.1962.10540810