INLS 490 - A Digital Gazetteer of North Carolina
Sam Aamot, Jiayi Xu, Anna Beckom
May 3, 2022

Cultural Heritage and Industries of North Carolina

Introduction

Our contributions for the NC Gazetteer project centered around North Carolina's rich cultural heritage. Expressing the state's cultural heritage took shape in three ways: accumulating lists of the state's musicians, notable filming locations, and the Department of Natural and Cultural Resource's Highway Historical Marker Program entries. Together, these entries and additions to the North Carolina Gazetteer enrich the project's sense of connections between the human and cultural histories linked to places. While by no means exhaustive, these datasets act as a starting point for further investigations into North Carolina's current and historic links to entertainment industries, artistic productions, and other cultural developments.

Part 1: NC Highway Historical Marker Program

CSV Documentation

The North Carolina Highway Historical Marker Program (NCHHMP) data have been separated into two .csv files. First, the aggregate dataset represents all 1,617 highway markers and available corresponding data. Second, a separate data subset was identified in order to further identify markers related to persons related to North Carolina's cultural industries, including writers and poets, musicians, and filmmakers and actors.

NCG_Historic_Highway_Markers

Aggregating the Historic Highway Markers data went through several stages. First, I visited the NCHHMP GIS page, and downloaded a .csv of 1,000 data points.[1] In addition, I queried Wikidata to check the dataset and generate QID's for each marker. I queried using the NC Highway Historical Marker Program ID (P9492), which generated 1,617 results. After recognizing that the dataset from the GIS page was incomplete, I emailed the NC DNCR and was sent a dropbox folder of ARCGIS Shapefiles, which Dr. Shaw was able to convert into .csv and .xlsx files for me to work with.

However, several discrepancies existed between the dataset I downloaded from the existing GIS map and the GIS shapefiles sent to me, resulting in the need to aggregate elements of the two datasets. Below includes a table of each dataset's incorporated data elements and a general description of how these data were defined.

[1] https://nc.maps.arcgis.com/apps/webappviewer/index.html?id=c53c699fdac146dbaccd215c986ac3d0

Data elements captured:
Note: the usage note and description are my interpretations of the data, as I have not been able to locate any documentation.

Field name (original GIS download [1,000 entries])	**Usage note and description**	**Field name: (updated GIS shapefiles [1,617 entries])**	**Usage note and description**
		ID	4-digit identifier, source unidentified
Marker ID	Alphanumeric identifier assigned by DNCR	Marker ID	Same as original data
		Prefix	Alphabetic character of Marker ID
		Suffix	Numeric character of Marker ID
Marker Title	Title of the marker	Marker Title	Same as original data
Marker Text	Text included on the marker	Marker Text	Same as original data
Location	Description of the physical location of the marker	Location	Same as original data
		County ID	Numeric values 1-100, assumed to match alphabetical order of NC's 100 counties
Sketch	Free-text description of the history associated with the marker, including references to textual sources and other Marker IDs	Sketch	Same as original data
Requestor	"Staff" or names of individuals assumed to have been involved in the creation of the marker	Requestor	Same as original data
		Notes	Free-text, information regarding planned or needed updates to the marker and other general notes on marker's condition

Year Cast	Year the marker was created	Year Cast	Same as original data
Years Replaced	Year or years the marker was replaced	Years Replaced	Same as original data
Main Term	Often matches the Marker Title	Main Term	Same as original data
Main Term Non Permutated	Unsure how this differs from the Main Term and Marker Title	Main Term No	Seems as though the title was cut off – seems to match "Main Term Non Permutated" from other dataset
		GPS	Boolean "Y" or "N" field
		Buffer	Numeric values 1-5, unknown reference
Longitude	Coordinate values	Longitude	Same as original data
Latitude	Coordinate values	Latitude	Same as original data
NCDOT District	Numeric values 1-14	DOT district	Same as original data

The main reason I needed to aggregate the datasets was because while the shapefiles contained entries for all existing 1,617 markers, the "sketch" values were erroneously truncated. Because the "sketch" field added important data, including references to textual sources and other historical highway markers, which I planned to extract, it was important to combine the full sketches from the smaller dataset into the larger one. I used the cell cross function in OpenRefine to do so, bringing full sketches from the smaller dataset into the larger one. I decided to move forward using the fuller dataset as my base dataset because it contained new, useful fields for other data capture.

Final dataset documentation: added fields

Field name	**Usage note and description**
ncg_id	Left blank, to be generated
ncv_type	Created "ncv:historicalMarker" type to express these locations.
reconciled_marker_id	This field was created in OpenRefine using the reconciliation service with Wikidata.
wikidata_id_nc_highway _historical_marker_program	This field was created in OpenRefine using the reconciled_marker_id field to generate the associated Wikidata QID for each marker.

associated_marker	This field was created in OpenRefine by identifying notated marker IDs from the sketch column. After attempting to reconcile these to generate Wikidata QIDs, I opted instead to leave the marker IDs as they were because reconciliation seemed to be causing errors and duplicates. These values could be reconciled to Wikidata in the future.
reference	This field was created in OpenRefine by extracting "references" from the "sketch" column. The original text of the "sketch" column, including references, was preserved but could be eliminated if deemed extraneous.
county	This field was created in OpenRefine by conducting a cell cross with a dataset I created linking numeric values 1-100 to an alphabetized list of North Carolina counties. While "county ID" was included in the larger GIS shapefile dataset, counties were not identified by name.
creator	Name of creator (myself)

NCG_Historic_Highway_Markers_subset

The subset I created focused on individuals identified in the NCHHMP dataset who are related to North Carolina's music, film, or literary history. While only a few of these figures are duplicated in the musicians dataset, I wanted to experiment with adding more context and unique identifiers for some of these individuals.

Final dataset documentation: added fields

Field name	**Usage note and description**
wikidata_id	Identifier for the person described by the historical marker
archives_at	URL to the finding aid for collections associated with the individual. Only included links to UNC Wilson Library's collections as I worked this past semester to create Wikidata pages for collection creators using the "archives at" Wikidata property.
Unique identifiers	Added fields for SNAC Ark IDs [https://snaccooperative.org/], FAST Linked Data IDs [http://experimental.worldcat.org/fast/], ISNI IDs [https://isni.org/page/search-database/], and NCPedia IDs [https://www.ncpedia.org/] in addition to LOC, VIAF, and WorldCat
place_of_birth	NCG ID for individual's birthplace
place_of_death	NCG ID for individual's place of death
date_of_birth	Numeric value for the year of the individual's birth. Chose to forgo date and month because these details are captured by unique identifiers.

date_of_death	Numeric value for the year of the individual's death. Chose to forgo date and month because these details are captured by unique identifiers.
qid_place_of_burial	Wikidata QIDs for names of cemeteries that may not be in NCG yet
associated_location	Names of places or other identified locations [universities, high schools] noted in "sketch" field. I kept this field because not all named places had a corresponding NCG ID, but I thought they might in the future.
associated_location_id	Wikidata QID or NCG ID
occupation	Identifier to categorize individuals to match other datasets
genre	Identifier to categorize individuals to match other datasets
contributor	Name for group member whose dataset these individuals relate to

Research Process

Sources

"NC Highway Historical Marker Program." NC DNCR. Accessed May 2, 2022. https://www.ncdcr.gov/about/history/division-historical-resources/nc-highway-historical-marker-program

Consolidated project page for the NCHHM program. Page includes links to the search interface (included below), background for the project, information on application processes for markers, damages, and contact information. The site also includes a link to the "Historical Marker Guide," a Word document identifying Marker Districts and corresponding counties, as well as more information about the program and all current markers. This was the main source for this project, and should be consulted for future data clean-up.

"North Carolina Highway Historical Markers." ArcGIS web application. NC DNCR. Accessed May 2, 2022. https://nc.maps.arcgis.com/apps/webappviewer/index.html?id=c53c699fdac146dbaccd215c986ac3d0

Source for the smaller dataset containing full sketches. This ARCGIS map contains pinned locations for most markers. Many of the pins also seem to include a link to an image for the marker, which could be used in the future. This page is published by the NC DNCR.

"North Carolina Highway Historical Marker Program - Search." North Carolina Office of Archives & History - Department of Cultural Resources. Accessed May 2, 2022. http://www.ncmarkers.com/search.aspx

Direct link to the search interface for the Highway Historical Marker Program. Search allows faceting by county, marker ID, and tags, which include categories such as "arts," "education," and other identifiers that were not available in any of the datasets shared with me or found online.

"NC Highway Historical Marker Program ID." Wikidata. Accessed May 2, 2022. https://www.wikidata.org/wiki/Property:P9492

Wikidata item page for the program's ID structure in Wikidata.

"North Carolina Highway Historical Marker Program." Wikidata. Accessed May 2, 2022. https://www.wikidata.org/wiki/Q21779079

Wikidata item page for the Historical Marker Program, including information about the program and how it has changed and been updated over time. This resource is important for finding updated information on the program and project, and is evidence that the program actively uses Wikidata as an additional tool.

"Search the Special Collections." The Louis Round Wilson Library Special Collections. UNC Chapel Hill Libraries, Accessed May 2, 2022. https://library.unc.edu/wilson/visit/search/

This source is listed as the search page for Wilson Library's special collections. It is referred to by collection in the dataset for any individuals having an "archives at" property/value pair in a corresponding Wikidata item.

"List of Counties in North Carolina." Wikipedia. Wikimedia Foundation, March 30, 2022. https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina

This source was used to create a list of all North Carolina counties in alphabetical order. This list was then used with cross cell functions in OpenRefine to match "county ID" to "county" column.

The Historical Marker Database. Accessed May 2, 2022. https://www.hmdb.org/

A source that was not used, but could be useful to explore in the future. It contains some historical markers related to North Carolina that are not part of the Highway Historical Markers program. The site apparently contains 3,620 North Carolina markers, some of which are part of the highway marker program. It also includes "war memorials," which might be useful for the gazetteer if any future groups choose to integrate data related to other commemorative sites.

Trials and errors, and other editing processes

I encountered several issues when working with the dataset that should be acknowledged and documented. First, when using OpenRefine to reconcile Marker IDs, it became apparent that some markers had duplicate Wikidata IDs. I decided to choose one Wikidata ID as it seemed the other option was to forgo reconciliation altogether. I manually reconciled over 100 items because they had not been matched in the first reconciliation attempt. Additionally, marker H-34 could not be reconciled, and I could not find it in Wikidata through a manual search.

Another instance of guess work includes the county IDs from the full dataset and the county name I created using cell cross in OpenRefine. I am assuming that the County ID (numeric values 1-100) corresponds to an alphabetical list of North Carolina's 100 counties, but I could not find this documented anywhere. The list seems to match the correct counties when compared to the marker's county in the search interface, but I did not check each county. I also did not have time to use cell cross to match NCG IDs for each county, so that column has been left with the county name as the only identifier.

Though this is addressed in the next section as well, it should be noted here that the highway marker data will change. I was notified that there will be updates to it in May 2022, and the nature of the markers is that they are replaced, eliminated, and added continuously. In addition to the markers themselves being moved and changed, some of the location description data in the dataset also indicates vague or unclear locations for events, burials, homes, or other information conveyed by the marker. Many of the markers have descriptions that include descriptions such as "Buried 2 miles NW," which indicates the undefined nature of the markers in relation to the things they describe. This information was kept in the full dataset, but was not integrated into coordinate locations in any way, though I think it could be of use in the future and should be kept associated with each marker.

Finally, I noticed textual errors in the data that I could not fix due to my limited time and ability with regular expressions. Mostly existing within the "sketch" column, some errors remain, including tab spacing to delimit paragraphs, some html (<i>) characters, and some remaining junk characters that were not cleaned in the first cleaning iteration. Finally, when converted to .csv I think some of the characters lost their formatting and became erroneous. While these do not affect much of the important data, like coordinate locations and counties, it will take extra labor to clean fully.

Ideas for Future Work

There were several avenues left unexplored and considerations to keep in mind related to the NCHHMP dataset. First, I was made aware through correspondence with the NC DNCR that markers are added, subtracted, and moved, and that the program "will be making a number of changes mid to late May [2022]" (from e-mail correspondence with Ansley Herring Wegner,

April 25, 2022). While integration of the data into the gazetteer is a good first step, it may be worth considering these changes and how the data could be updated in the future.

Additionally, if there were more time I would have followed several other avenues with the data. First, the subset could be expanded to all individuals, and the markers could be further identified by what they commemorate [event, person, group, structure, organization]. Then, a standardized list of predicate identifiers could be added to describe each type of entity being commemorated. Another addition to the dataset could be more "archives at" URL identifiers for markers commemorating individuals. As a graduate assistant, I worked with Anna Goslen, Metadata Librarian, to create Wikidata items for over 1,000 collection creators in Wilson Library's special collections. I would assume that a percentage of the individuals commemorated in the highway markers program have archival collections in Wilson Library, which could be queried using the "archives at – Louis Round Wilson Library" and matched to pages for these collection creators. Finally, I decided to focus on including as many unique identifiers as I could think of for each individual, and I was wondering if there was a way for this process to be more automated through a query. I think it would be helpful to be able to generate these identifiers automatically for different subsets rather than relying on selecting them individually.

Finally, I think the NCG project could create a Wikidata project page in order to express and document all the ways it works with Wikidata. For the Wilson Library "archives at" project, I created a Wikidata project page, which notates documentation and property/value pairs, and gives others a sense of your decision making process and can act as a guide for other or future projects.

Part 2: North Carolina Musicians [Jiayi Xu]

CSV Documentation
Final dataset documentation:

Field name	**Usage note and description**
name	Name of the musician, musical group, and ensemble
wikidata_id	Wikidata identifier for the individual or group in the North Carolina music industry
type	This field was created in OpenRefine by adding columns from the reconciled column to generate the associated occupation/instance of each individual or group.
originate_from	This field was created in OpenRefine by adding columns from the reconciled column to generate the associated place of birth/ location of formation for each individual or group.

NCG_ID	NCG ID associated with each place that each individual or group originated from
Period (start)	This field was created in OpenRefine by adding columns from the reconciled column to generate the associated work period (start) for each group or date of birth for each individual.
Period (end)	This field was created in OpenRefine by adding columns from the reconciled column to generate the associated work period (end) for each group or date of death for each individual (if applicable).
Genre	Genre of musical works
Description	Abstract for each individual or musical group

Research Process

Sources Tools

Sources: https://en.wikipedia.org/wiki/Category:Musicians_from_North_Carolina
When I did research, I figured out that there is little aggregated information on the Internet. Thus, I mainly used Wikipedia and Wikidata as my sources and query them respectively through https://query.wikidata.org/ and DBpedia https://dbpedia.org/snorql/.

After I got a list of names and the abstract, I saved the result as JSON files. To display the result and clean data, I implemented OpenRefine and transformed cells by using the following command:

value.replace(/^http:\/\/dbpedia.org\/resource\//, "")
value.replace("_", " ")
value.replace(/ \(([^\)]+)\)/,"")

These allowed me to get rid of useless characters and get a list of names that could be reconciled directly. After using "human" as an entity and reconciling the "name" cell, I got the wikidata_id. I was then able to add new columns such as "genres", "Period (Start)", and "Period (End)" from reconciled values. For those values that were incomplete, I did research on the Internet, read their stories, and manually completed the list. Later on, I linked the place name with the North Carolina Gazetteer Identifiers by using the cross function.

Since the *Musicians from North Carolina* page contains 36 subcategories, I respectively create 36 CSV files. To aggregate these lists, I employed Conditional Formatting in Excel to highlight duplicate values.

Trial and error/other processes
Excel converts UTF-8 CSVs to ANSI. As a result, some of the characters in my list will not display correctly. I saved the file as CSV UTF-8 to avoid this from occurring.

Furthermore, when I open my CSV in Excel, it formats the date automatically, which causes some errors. I manually changed it to string type to avoid formatting.

Ideas for Future Work
Some of the sources are inconsistent and outdated when conducting research. For example, many musicians such as George Huntley do not have a specific date of birth. Thus, future work concerns a deeper sources analysis.

Additionally, when doing research, I discovered that Chapel Hill is home to a number of indie rock bands. As a result, visualizing the data on maps and labeling different genres would be intriguing. We can learn more about how music genres are influenced by geography, history, and culture.

Finally, many of the musicians on my list have several occupations, including but not limited to actors, veterans, and politicians. It would be interesting to investigate the connections between music and other industries.

Part 3: North Carolina Films

CSV Documentation

Final dataset documentation:

Field name	**Usage note and description**
film_name	Name of the film (includes the categories of “Film” and “TV Film” but not “TV Series”)
wikidata_QID	Wikidata ID associated with each unique film
IMDb_ID	IMDb ID associated with each unique film
IMDb_URL	The direct URL to each unique film on IMDb’s website. The URL is composed of the prefix “https://www.imdb.com/title/” with the IMDb ID

	added to the end; so, each URL is "https://www.imdb.com/title/**IMDb_ID**"
IMDb_locations_URL	The direct URL to each unique film's filming location page on IMDb's website. The URL is composed of the prefix "https://www.imdb.com/title/**IMDb_ID**" plus "/locations?ref_=tt_ql_sm" added to the end; so, each URL is "https://www.imdb.com/title/**IMDb_ID**/locations?ref_=tt_ql_sm"
place	The location(s) each film was filmed in
ncg_ID	NCG ID associated with each place (filming location)

Research Process

Sources & Tools

"Category:Films shot in North Carolina" Wikipedia. Accessed May 2, 2022
https://en.wikipedia.org/wiki/Category:Films_shot_in_North_Carolina

To compile this list in an easier manner I used dbpedia.org/snorql/ with this query below and then copy and pasted the results in excel
SELECT ?film WHERE {
?film <http://purl.org/dc/terms/subject>
<http://dbpedia.org/resource/Category:Films_shot_in_North_Carolina> .
}

North Carolina Film Office. Accessed May 2, 2022
https://www.filmnc.com/1950-1979

This website includes a "Film & Television Archive" with six individual pages, separating films filmed in North Carolina by date. The dates separated are as follows:
1950-1979
1980-1989
1990-1999
2000-2009
2010-2019
2020-Present

To compile this list in the easiest manner I copy and pasted film names from each page into excel

https://query.wikidata.org/#%23Cats%0ASELECT%20%3Fitem%20%3FitemLabel%20%0AWHERE%20%0A%7B%0A%20%20%3Fitem%20wdt%3AP915%20wd%3AQ1454.%20%23%20Must%20be%20of%20a%20cat%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22%5BAUTO_LANGUAGE%5D%2Cen%22.%20%7D%20%23%20Helps%20get%20the%20label%20in%20your%20language%2C%20if%20not%2C%20then%20en%20language%0A%7D

This link directs to the Wikidata Query Service with the specific query I used to generate a list of "items" (QID from Wikidata) and "itemLabel" (film title). I then copy and pasted all itemLabels into excel.

Trials and errors, and other editing processes

I could not find a specific data set for all films filmed in North Carolina. So, to compile the most exhaustive list I used the three sources and tools listed above to make my complete dataset. After pasting each individual list into excel I used functions to clean up the film titles to be the same characters and remove duplicates. This left 368 unique films.

I then uploaded this excel file into OpenRefine, reconciled the film_name columns, and manually checked anything that did not automatically match. From these reconciled values I created the new columns of wikidata_QID and IMDb_ID. By utilizing the join columns feature I then created the columns of IMDb_URL and IMDb_locations_URL.

Next, I needed to scrape the data from IMDb. I used the IMDb_locations URL column to fetch the raw HTML data with a throttle delay of 5000 ms (this took about two hours, so in the future, I know to plan accordingly for the time required for this process). To parse the result into a new column called "place" with the GREL expression:

```
filter(
  forEach(
    value.parseHtml().select("dt"),
    node,
    node.htmlText()
  ),
  place,
  place.endsWith("North Carolina, USA")
).join("|")
```

Next, I split any multi-valued cells so I could have each place be its own record. To clean up the data in the place column I used the GREL expression: value.match(/(.*), North Carolina, USA/)[0] along with the manual work by hand.

Lastly, I added the NCG_id column with a GREL cross function from the NCG csv file.

Ideas for Future Work

The dataset, or list of films filmed in NC I complied, includes films produced around 1950 and onward. So, it could be interesting to dive into films that were filmed in NC prior to this time. Additionally, I only included "Films" and "TV Films" in my dataset, and many "TV Series" were filmed in NC that could be included in the Gazeteer.

I think it could also be very interesting to explore/add filming cities in the most popular places based on filming locations. For example, a high amount of filming takes place in the NCG place, Wilmington. I think it would be interesting to see if there are any historic markers, landscapes, etc specific to Wilmington that could be linked to the filming location. To do this, one would need to find detailed sources/archives of Wilmington relating to filming. Furthermore, many films took place at Universities in North Carolina which do not have their own NCG IDs, so this could be added/linked.

Lastly, as films are produced this dataset could constantly be updated.

Sample RDF Documentation

The following are some examples for possible RDF triples that could be created from the dataset. I have followed documentation for Wikidata properties like “archives at” and “place of birth” or “place of burial” to create interoperability when reconciling or using Wikidata in the future. Additionally, the Historical Marker Program ID in Wikidata identifies each marker as “instance of – commemorative plaque,” which is why the predicate for the first triple is “commemorates.” The last example is a sample RDF for a triple included in the films cultural industry topic.

[Q111464005] (QID for historical marker) – (commemorates) – [Q109612] (Thelonious Monk QID)

[Q109612] (Thelonious Monk QID) – (birthplace) – [NCG13021] (NCG ID for Rocky Mount)

[Q109612] (Thelonious Monk QID) – (associated with or collaborated with) – [Q7346] (QID for John Coltrane)

[Q109612] (Thelonious Monk QID) – (genre) – jazz

[Q18538790] (10 Rules for Sleeping Around QID) – (filmed in) – [NCG02916] (NCG ID for Charlotte)

Credits & acknowledgements

Project group members
Sam Aamot – contributed to data related to the NC Highway Historical Marker Program. Cleaned and contributed to the generation of 1,617 historical markers and identified 17 individuals to generate in-depth biographical descriptions, focusing heavily on unique identifiers and related places. Also relied on correspondence with Ansley Wegner, Administrator, North

Carolina Highway Historical Marker Program and Andrew Edmonds, GIS Technical Support Analyst with the State Historic Preservation Office.

Anna Beckom – contributed data related to films and TV films filmed in North Carolina. Cleaned and contributed 368 film titles, and respective URL's to IMDb (IMDb is an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, production crew, and personal biographies, plot summaries, trivia, ratings, and fan, and critical reviews). The csv file attached includes 756 entries of places with most having a respective NCG ID.

Jiayi Xu - contributed data related to musicians or musical groups in North Carolina. Cleaned and contributed 1228 individuals or musical groups to generate descriptions, focusing heavily on unique identifiers and related places.