Geocoding entities in Digital Gazetteer of North Carolina with Wikidata and GeoNames

The main goal of this project was two-fold. The first goal was to reconcile entities in the Digital Gazetteer of North Carolina (hereafter the Digital NCG) to their corresponding entities in Wikidata and GeoNames Knowledge Bases (KB) within OpenRefine to facilitate data extraction from these two sources. Furthermore, Wikidata and GeoNames have been widely used data hubs for linked open data. Therefore, linking to them is beneficial for increasing the Digital NCG's visibility in the Linked Open Data universe and will assist in future interlinking with other datasets as an open and freely accessible resource.

The second aim was to explore Wikidata and GeoNames as sources for obtaining longitude and latitude values to map and visualize in Powell's *Gazetteer* in GIS. For historians accustomed to working with imperfect, fragmented, and often imprecise, and conflicting information about places, the absolute spatial accuracy of locations may not be their primary concern in answering their research questions. Visualization in GIS cannot replace or substitute the value of Powell's textual descriptions. Mostern summarizes the unique value of gazetteers as "gazetteers make it feasible to record uncertainty, textual references, multiple perspectives and temporal change,"[1] which are difficult to represent in GIS. However, knowing where things took place helps us reconstructing the past. GIS visualization has a way of highlighting or bringing out aspects that can go unnoticed when examining textual descriptions alone. GIS-enabled historiography or Historical GIS (HGIS) can give us new insight and deepen our understanding of places.

This project does not enrich the Gazetteer by adding further descriptive, textual information directly nor augmenting the number of entries. However, the JSON outputs retrieved from GeoNames API contain a wealth of information such as population, administrative division, hierarchy, elevation, type of place, point coordinates, or bounding box coordinates. Similarly, entities that are now reconciled to Wikidata can harvest additional data from Wikidata in the future. For this project, I have only extracted four elements from the GeoNames JSON output (11654 entities) and reconciled Wikidata results (3791 entities): longitude, latitude, GeoNames ID, and Wikidata ID.

Wikidata & GeoNames

Wikidata is a cross-domain, community-created Knowledge base with varying degrees of completeness and accuracy. GeoNames is a knowledge base of geographical locations primarily built on official public sources,[2] such as national mapping agencies the United States National Geospatial-Intelligence Agency. Nevertheless, it also allows user contribution. Both Wikidata and GeoNames provide API services for reconciliation and data retrieval. For reconciling

[1] Ruth Mostern, "Historical Gazetteers: An Experiential Perspective, with Examples from Chinese History," *Historical Methods* 41, no. 1 (2008).
[2] "GeoNames Data Sources," accessed May 12, 2021, https://www.geonames.org/datasources/.

entities in the Gazetteer to Wikidata, this project used the built-in reconciliation service in OpenRefine.

Wikidata was only used at the beginning of the project. The process nevertheless revealed the advantages and disadvantages of using Wikidata reconciliation service within OpenRefine. It also offered an opportunity to evaluate Wikidata as a data source. One advantage was that the reconciliation service is already integrated into the OpenRefine platform; it does not require any additional setup, unlike with GeoNames. However, high ambiguity resulting from incomplete records, unevenness of coverage concerning different geographic feature types, and the entities of the same feature types, presented difficulties in linking entities in the Digital NCG to Wikidata. Some feature types such as mountains, counties, towns are covered much better than others, for instance, creeks and branches. Reviewing the matched results was a particularly time-consuming process, as suggested matches do not always display sufficient information in preview mode for disambiguation; thus, possible matches had to be verified by viewing them in Wikidata using hyperlinks provided.

Earlier on, it became clear that GeoNames is a more suitable dataset for the entity linking *Gazetteer* entries as it contains more matching entities than Wikidata. Wikidata covers multi-domain, with currently over 96 million records.[3] GeoNames inlcudes over 25 million geographical names;[4] therefore is a more focused dataset for this project. Moreover, it also proved to be much faster in matching entities. About 90% of reconciliation was done using GeoNames; however, through reconciling against GeoNames, it was also possible to obtain Wikidata Q ids from GeoNames JSON outputs when present.

The main challenge with the entity matching process for both datasets was variant place names existing for a single geographical place. Powell's *Gazetteer* contains many variant forms of place names. Both Wikidata and GeoNames were unable to match most of them. The matching algorithm of Wikidata is based on the Damerau-Levenshtein distance of two or edit distance of two, which refers to the number of single-character transformations, such as deletion, insertion, and substitution required to produce identical strings.[5] If the place name in the Gazetteer differs from those of Wikidata entries by more than two single-character edits, Wikidata fails to return a match. Partial matches do not work either in both. GeoNames showed a similar level of matching capability when it comes to variant names. However, in GeoNames, when it lists variant names for entities, it can match up regardless of the types of variants. However, its list of variant names is nowhere near exhaustive.

Over 700 such variants have been detected so far. The most common types of variations observed (Powell's in italics) appear to be misspellings (e.g., *Barbecue* vs. Barbeque), omissions (e.g., *Arington Branch* vs. Arrington Branch), additions (e.g., *Bob's Knob* vs. Bob Knob, *Blues Pond* vs. Blue Pond), substitution (e.g., *Lizzy* vs. Lizzie), abbreviations and spacing of compound

[3] "Wikidata:Statistics," updated January 24, 2021, accessed May 14, 2021, https://www.wikidata.org/wiki/Wikidata:Statistics.
[4] "About GeoNames," accessed May 13, 2021, https://www.geonames.org/about.html.
[5] "Query string query," 2021, accessed April 8, 2021, https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html#query-string-fuzziness.

names (e.g., *Beargrass* vs. Bear Grass). Transliteration of native Indian place names (e.g., *Ocanecchi* vs. Occoneechee) was another variant name source. Many places in the gazetteers now have been renamed, and these also caused the matching failure. Places with offensive and racially sensitive names have long been replaced and disappeared from the public domain, but Powell's Gazetteers published in 1968 and the revised second edition published in 2009 still retrained many such historic and colloquial names.

Disambiguation of entities with the same or similar names was also a time-consuming process, even with GeoNames. There are at least 110 entities called "Long Branch" in GeoNames for North Carolina alone. It is not uncommon to find multiple places with the same name within the same county; they all have distinctive geocoordinates, unique IDs, and are not adjacent to each other. There are five Long Branches in Haywood county, six in Jackson county, and likewise, the same pattern was observed in other counties and other place name entities. These had to be verified and matched one by one by consulting a map and Powell's description.

Creeks, rivers, and streams presented slightly different kinds of matching issues. In the Digital NCG, the organization of entities is primarily by counties. Creeks, rivers, and streams could rise in one county but often cross the county boundaries and flow into another county. In many cases, trying to reconcile them based on the county mentioned in Powell's Gazetteers did not initially find the match. They were often listed under another county in GeoNames. These also had to be verified by reading Powell's description as well as looking up the map.

Another possible cause of the matching failure could be that some of the entities in Digital NCG have been assigned the incorrect feature types. For instance, not every entity that contains the word "Mountain" in its title is an actual mountain; it could be a name for a town. Likewise, a placename having "beach" at the end may not be referring to a beach in a geographic sense; it might be a name given to a populated place. In some cases, Powell's description is not explicit enough to make a positive verification. However, in other cases, some entities in the Digital Gazetteers have been assigned incorrect feature types, primarily based on the name. Also, places that share a common name may not belong to the same feature type. Same place entities could have been labeled differently in different data sources.

Furthermore, feature types in the Digital NCG do not always align with feature classes and feature codes in GeoNames or classes and subclasses in Wikidata. These differences could have hampered the matching process. In a similar vein, reconciling the entities in the NCG feature type "Community" was also problematic. It may be too under-specific, containing many types of places; matching against GeoNames only yielded 2264 matches out of 3599 entities present in the *Gazetteer*. Many in this Community category could be better placed under other feature types.

My method of systematically working through feature types by counties may have neglected to include some related feature codes for the GeoNames dataset. County boundaries would have changed over the years, resulting in some entities belonging to different counties than recorded in the *Gazetteer*. A closer reexamination of the entries and Powell's description in the future may result in more matches to be found.

Methodology

Configuring the GeoNames reconciliation service in OpenRefine was not straightforward. After trying out three different methods, it was decided to connect to GeoNames API directly to get the data. It turned out to be a better approach as it allowed direct programmatic access to the gazetteer dataset and the flexibility to craft more specific queries. Retrieving data from GeoNames API with OpenRefine was based on a GitHub tutorial.[6] The steps to constructing the query URL are well outlined in the tutorial, and there exist multiple resources describing step-by-step processes of retrieving data from GeoNames API. Therefore, it is not replicated in this report. The tutorial was adapted to build a custom search request URL for the Gazetteer dataset.

Once the query URL was configured, fetching data from GeoNames API is done using the drop-down menu at the label column heading and selecting 'Edit column → Add column by fetching URLs.' The basic pattern of GREL expression used for creating a search request URL follows the pattern below:

"http://api.geonames.org/searchJSON?q="+escape(value, "url")+"&adminCode2=[FIPS number for each individual counties] &adminCode1=NC&country=US&featureClass=[feature class]&featurecode=[feature code]&maxRows=1&username=[GeoNames username]&style=full"

The adminCode 1 refers to the first-order administrative division, a state, for the US; the adminCode2 is for the second-order administrative division, counties within each state. Different combinations of GeoNames feature classes, feature codes, and the US coding standards (FIPS numeric codes) for North Carolina counties were used for reconciliation. Specifying the county, feature class, and feature code allowed quite precise searches to be carried out, reducing the result reviewing time. The GeoNames feature class and feature code combinations use for the project are listed in the table below. The complete list, which contains some 645 feature codes in 9 feature classes, can be found here: https://www.geonames.org/export/codes.html

Mapping NCG feature types to GeoNames Feature Classes and Feature Codes

NCG feature types	GeoNames Feature Class	Feature Codes
nct:Community (3599)	P	PPL
nct:Creek (2258); nct:Branch (1416)	H	STM
nct:Mountain (921); nct:Knob (465); nct:Bald (94); nct:MountainFeature (77); nct:Top (91); nct:Mountains (34); nct:Rock (17); nct:Elevation (27); nct:Hill (24)	T	MT, RDGE, CLF

[6] "Tutorial-Geocoding-Crash-Course," updated December 10, 2018, 2018, accessed April 13, 2021, https://github.com/mapninja/Tutorial-Geocoding-Crash-Course.

nct:Township (819)	A	ADMD, ADM3
nct:Gap (634)	T	GAP
nct:Town (531)	P	PPL
nct:Ridge (509)	T	RDGE
nct:Swamp (286)	H	STM, SWMP
nct:FormerCommunity (262)	P	PPL, PPLH, PPLW
nct:Cove (205)	T	VAL
nct:Lake (187); nct:Millpond (37)	H	LK, RSV, DAM, SWMP
nct:Point (187)	T	CAPE, MT
nct:Island (180); nct:Islands (17)	T	ISL
nct:Fork (153); nct:Prong (30); nct:Run (67); nct:Stream (6)	H	STM
nct:Bay (143); nct:Sound (18)	H	BAY, SWMP
nct:River (134)	H	STM, WTRC
nct:Pond (130)	H	LK, RSV, SWMP
nct:County (127)	A	ADM2, ADMD
nct:Valley (106)	T	VAL
nct:Falls (80)	H	OVF
nct:Area (78)	T, P	PPL, PLN, VAL, RDGE, ISL, CLF, DPR, AREA
nct:PopulatedPlace (54)	P	PPL
nct:Pocosin (49)	H	SWMP
nct:Inlet (40); nct:Channel (24); nct:Canal (22)	H	CHN, INLT
nct:City (37)	P	PPLA2
nct:Beach (20)	T	BCH
nct:Shoal (20)	T	BAR
nct:Forest (18)	V	FRST
nct:Spring (18)	H	SPNG
nct:Gut (14)	H	SMT, INLT
nct:Bend (12)	H	STMB

Contents of the CSV files

I have added to the dataset the following new columns: 1) geonames_id_number, 2) geonames_json output, 3) latitude, 4) longitude, 5) geonames_id (URL) columns. I have also filled in the existing wiki_id column whenever Wikidata IDs were present in the GeoNames JSON output to allow data extraction from Wikidata in the future. Many place-related records in Wikidata also contain GeoNames Ids.

All the entities that were initially geocoded using Wikidata were also found in GeoNames. So far, 11654 out of 16812 entities have been geocoded due to time constraints. A small proportion of entities listed in Powell no longer exist or were planned and proposed but never came into being. For these entities, there are no corresponding records in datasets like GeoNames and

Wikidata. However, it is feasible to complete the geocoding for all entities using georeferenced historical maps.

The second CSV file is a stand-alone file, not necessarily intended to be incorporated into the main file; it is a list of name variants. Column B contains place names as appeared in Powell's Gazetteer, and Column C has their equivalents in GeoNames or Wikidata. The list is mainly for reference purposes only. They highlight the sort of spelling variants present in the source and places renamed in the past. Tracing name changes and variant names, in itself, could be developed into a project of its own for creating a comprehensive list of alternative place names in North Carolina.

Visualizing in GIS

Graphics can reveal what statistics might miss, and patterns in historical relations that we were aware of become even more pronounced when we can visualize it. Gaps in data, missing data types can be picked up. Furthermore, GIS is not just a visualization tool; it can also sort, integrate, and aggregate complex geographic location data with spatial attributes, making it easier to analyze demographic changes, migration and settlement patterns, and urban development. It also allows us to explore a particular geographic feature type in isolation. Data visualization helps detect errors in geocoding, spot unusual distribution and outliers, and evaluating the output.

The initial visualization attempt was made in ArcGIS online; currently, every entity is drawn as a point coordinates. However, GeoNames JSON output did provide coordinates of the bounding box (north, south, east, and west) for certain feature types. It would be possible to use bounding box coordinate or polygon boundaries instead of point coordinates in mapping entities like townships, towns, and counties. Incorporating administrative boundary polygons into the map could be helpful, although we must bear in mind that these have changed over time.

Integrating historical maps as raster layers will enable us to locate places that no longer exist. The current GIS visualization is static, but interactive and dynamic visualization is also possible, as in the *Map of Early and Modern London* (MoEML) project.[7] We could also add animations to show various changes that happened in North Carolina.

Future Direction

The initial plan for this project was to finish the Geocoding, convert the data into GeoJSON, and subsequently to Linked Places Format. However, the entity matching process alone turned out to be much more laborious and time-consuming than had initially expected; many entities with variant names that were not matched automatically had to be handled manually.

[7] https://mapoflondon.uvic.ca/index.htm

Disambiguation of entities often required looking up the map and verify one by one. As the project ended up being incomplete, the next step for this project would be geocoding the remaining 5000 entries, then proceed with converting data into Linked Places Format.

There are many other possible future projects for the Digital NCG. Converting the current version into a temporally enabled gazetteer would be particularly useful to historians. Temporal data could be incorporated as an attribute, such as the inception of a town, when the country boundary changed, closure of mines or mills. Using Powell's descriptions, it is possible to trace changes that happened to places over time and their timelines for some locations. If the details are not in Powell's *Gazetteer*, future class students can research and add them to the Digital NCG. Timeline visualization would make it possible to chart out the settlement patterns of towns, townships, villages, how the landscape of North Carolina changed over time, and how these changes affected human activities.

Similarly, adding event attributes to the Digital NCG is another possible direction it could take, making it a historical event gazetteer. The concept of location is not as significant to historians as that of place. Mostern and Johnson assert that "Location become places only when events occur that cause them to become imbued with meaning,"[8] highlighting the importance of relationships between time, location, and event in historical gazetteers.

Another possible future project that piqued my interest while going through the Gazetteer entries is researching former communities that no longer exist physically, for instance, towns, old mills, and dams now submerged under Lake Norman and Lake Fontana. At least ten former communities in Swain County listed in Powell's are now under the waters of Fontana Lake.[9] They are not well-covered by other existing datasets, and it was not possible to get coordinates for these places using GeoNames or Wikidata. Such a project could add distinctive value to the Digital NCG and a unique contribution to the history of North Carolina. In a similar vein, enriching data on elusive and imaginary places that Powell had taken from old maps, counties, and towns that were planned and proposed but never came into being, is another area that could add significance to the Gazetteer. Study of abandoned villages and towns, too, could reveal various aspects of the economic and social history of North Carolina.

A Digital Gazetteer project such as this has so much potential for being more than mere indexes of names places presented in a digitally enhanced format. The real value of this Digital NCG will be its ability to allow and provoke historical reasoning for historians and students.

---

[8] R. Mostern and I. Johnson, "From named place to naming event: creating gazetteers for history," *International Journal of Geographical Information Science* 22, no. 10 (2008/10/01 2008): 1092, https://doi.org/10.1080/13658810701851438, https://doi.org/10.1080/13658810701851438.

[9] These are Bushnell NCG02255, Collinwood NCG03373, Dorsey NCG04390, Ecola NCG04745, Forney NCG05425, Fry NCG05608, Judson NCG08029, Marcus NCG09581, Noland NCG10872, and Ritter NCG12840.

Bibliography

Elasticsearch. "Query string query." 2021, accessed April 8, 2021, https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html#query-string-fuzziness.

GeoNames. "About GeoNames." accessed May 13, 2021, https://www.geonames.org/about.html.

———. "GeoNames Data Sources." accessed May 12, 2021, https://www.geonames.org/datasources/.

Maples, Stace. "Tutorial-Geocoding-Crash-Course." Updated December 10, 2018, 2018, accessed April 13, 2021, https://github.com/mapninja/Tutorial-Geocoding-Crash-Course.

Mostern, R., and I. Johnson. "From named place to naming event: creating gazetteers for history." *International Journal of Geographical Information Science* 22, no. 10 (2008/10/01 2008): 1091-108.

Mostern, Ruth. "Historical Gazetteers: An Experiential Perspective, with Examples from Chinese History." [In English]. *Historical Methods* 41, no. 1 (2008): 39-46.

Wikidata. "Wikidata:Statistics." Updated January 24, 2021, accessed May 14, 2021, https://www.wikidata.org/wiki/Wikidata:Statistics.