Enhancing and Validating GeoNames Data with Digital Nautical Charts Data: A Case Study in the Mapping of Freeform Map Labels

Dakotah D. Maguire* [a], Jason C. Kaufman[a], Alexandre Sorokine[a], Robert Stewart[a]

[a] Oak Ridge National Laboratory, Oak Ridge, Unites States
* maguiredd@ornl.gov

Keywords: Digital Nautical Charts, GeoNames, open source, text similarity

Introduction

GeoNames (2022) is a substantial and evolving gazetteer with over 27 million geographic names and, with 150 million daily requests (About GeoNames, 2022), is widely used. As with any dataset that relies on community contributions, GeoNames users are commonly interested in extending, enhancing, and validating these data, often by comparing with exogenous data sources such as Open Street Map. Digital Nautical Charts (DNC) (National Geospatial-Intelligence Agency [NGA], 2022) is a substantial, long-lived, worldwide vector chart database for ship navigation. It is developed and maintained by the NGA, the charting authority for the United States, together with the National Oceanic and Atmospheric Administration (NOAA, 2022). Although modern updates are underway (e.g., S-57, S-100) (International Hydrographic Organization, 2018, 2000), there is a wealth of DNC data that continues to be updated and used in mapping services. Assessment of these two repositories for enhancing or extending one another is noticeably missing from the open literature. The opportunity is engaged here by raising the following question: can DNC data enhance GeoNames data by supporting validation, expanding aliases, and filling in missing data?

There are several challenges in answering this question. First, DNC geographic names are scattered across 144 feature classes (e.g., BH140 Rivers and AL015 Buildings) as freeform text in notes and text fields. These fields serve a wide array of functions, and text found therein may contain place names as well as other relevant data. However, a particularly promising feature class is Earth Cover Text (ECRText) data. ECRText is used in contextual labelling of named places (e.g., Atlantic Ocean); generic features (e.g., sandbars); and other information (e.g., Unexploded Ordinance). Second, placement of ECRText is determined by cartographic labeling practices and strongly depends upon local context. For example, some labels may be shifted to avoid obscuring underlying maritime features; geolocation is therefore only approximate, and the geometry amounts to the rectangular polygons of the text placement on the map sheet. Third, there is the potential for duplication of label information owing to the presence of four spatial scales: Harbour, Approach, Coastal, and General. Labels of the same object (e.g., Chesapeake Bay) may occur in all four scales. Resolving these duplicates would require a significant conflation effort, and we postpone this for future research.

GeoNames dataset is structured as a flat table with each row containing the geographic name itself with a latitude–longitude pair representing its coordinate and some additional information (e.g., one of nearly 700 feature classes). It is easy to interpret a particular location for a point-like feature (e.g., lighthouse or spring), but positions of GeoNames representing linear (e.g., rivers, beaches) or areal (e.g., administrative units, bays) features are also inexact. For example, in some cases, multiple, duplicated GeoNames points can be found along the same river.

The challenge we engage here is to detect instances of ECRText within GeoNames through textual similarity and spatial proximity to help determine where DNC can contribute new and supplementary information. Combination of these factors is commonly used in studies involving both geodata and unstructured text (e.g., Kim et al., 2017; Šimbera, et al., 2021). In this paper, we address this challenge by forging a linking capability between GeoNames and ECRText based on proximity and similarity. We describe the workflow and apply it to US waters. We find that of 14,224 ECRText instances, we can identify up to 1,103 ECRText features that could extend GeoNames as new named locations or new aliases, 230 ECRText features that could extend GeoNames with new aliases, and 12,439 ECRText features that could validate existing GeoNames locations.

Materials and Methods

Our region of interest is DNC Region 17 (DNC17), containing 14,224 ECRText objects and 222,772 GeoNames locations along US East Coast from 42° north to 33° north latitude (Figure 1). The DNC17 region is made up of 76 libraries (charts), each with its own set of ECRText features.

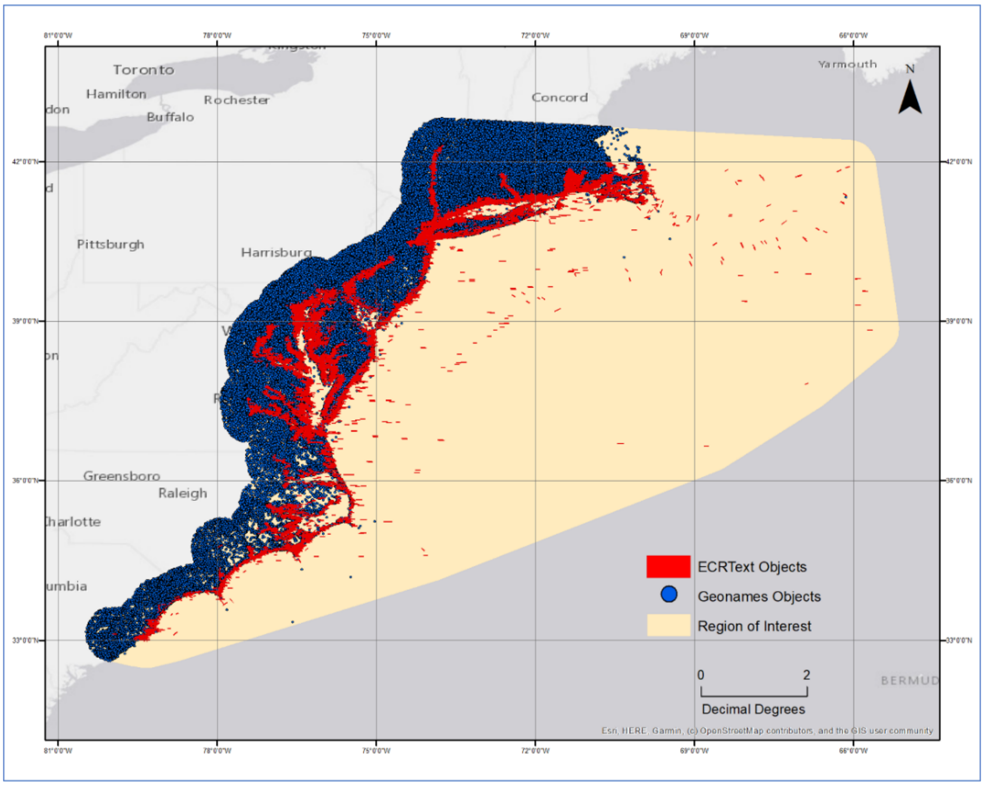

Figure 1: DNC Region 17 containing 14,224 ECR Text and 208,203 GeoNames.

We can compute the distance between any ECRText–GeoNames data pair as the closest distance between the GeoNames point and the ECRText bounding polygon. We define a pair as "sufficiently close" if their distance is less than a distance threshold d_t. To calculate the similarity of the names, we use a computationally effective trigrams method that results in the similarity values in the range of 0–1 (Dunn, 2020). We define a pair as "sufficiently similar" if the similarity value is greater than a threshold s_t. Using these two factors, we aim is to classify ECRText objects into one of three categories:

- **Confirming:** Here, an ECRText is part of a sufficiently close pair (distance $\leq d_t$) with an exact text–name match (similarity = 1). The interpretation here is that ECRText very close to a GeoNames location with exactly the same name provides some confirmation that the GeoNames is likely accurate.
- **Alias:** Here, an ECRText is part of a sufficiently close pair (distance $\leq d_t$) with a sufficiently similar name ($s_t \leq$ similarity < 1). The interpretation here is that ECRText found very close to a GeoNames location with nearly the same name may well be a new alias (i.e., alternate name) for the already existing GeoNames location.
- **New:** Here, an ECRText is either too distant from any GeoNames (distance $> d_t$) or is part of a pair with significantly different text–name matches (similarity $< s_t$). The interpretation here is that the ECRText object is too distant or too different to be reasonably associated with an existing GeoNames location. These may well be candidates for new locations to enhance the GeoNames dataset.
- **Discard:** Here, an ECRText is too distant from a GeoNames location with a label that is not associated with a geographic name (e.g., Unexploded Ordinance). The interpretation here is these place names are highly unlikely to have any clear benefit to GeoNames.

A semiautomated workflow (Figure 2) was developed to classify ECRText. ECRText is first cleaned and normalized using US Chart No. 1 (NGA, 2019). ECRText and GeoNames within d_t distance are considered matched pairs and unmatched otherwise. Matches are then compared for text similarities (s) and classified as confirming or alias. Unmatched pairs are manually reviewed to discard any results that are not obviously geographic names. The rest are candidates for new locations.

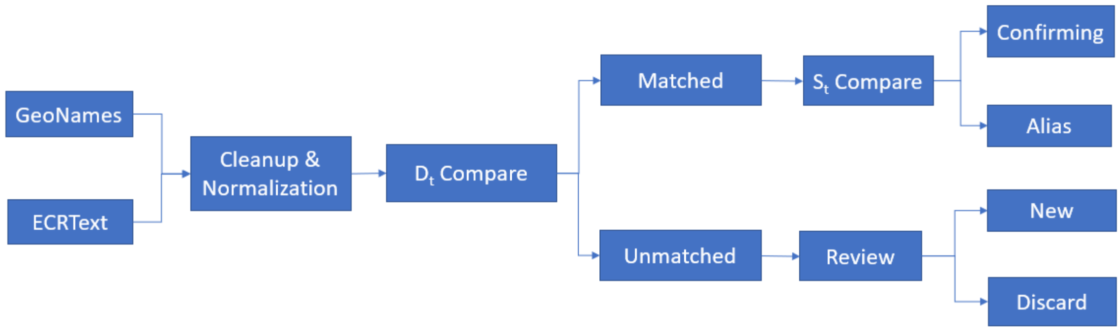

Figure 2: Semi-automated ECRText classification workflow.

Based on preliminary testing, a similarity threshold value of 0.8° and distance threshold value of 0.5° was found to provide reasonable classification outcomes.

Results

Using a buffered concave hull around ECRText features in the DNC17 region, we identified 14,224 ECRText objects and 208,203 GeoNames to use in the study area (Figure 1). Applying the workflow to these data yielded the classification results in Table 1.

Classification	Total ECRText Objects
Confirming	12,439 (87%)
Alias	275 (2%)
New	1,103 (8%)
Discard	407 (3%)

Table 1: ECRText object classification results.

The results demonstrate the significant value of DNC ECRText data to validate and enhance GeoNames data. The most significant benefit is validation with 87% of ECRText data classified as confirming. ECRText in the Alias (2%) and New (8%) categories has the potential to enhance and extend existing locations.

The workflow (Figure 2) represents an initial, successful semiautomated process, but opportunities for improvement exist particularly in the case of duplication. Examples include ECRText duplication over multiple scales, as previously mentioned, but potential duplication in merging multiple libraries may also require conflation workflows. Resolving these duplications would require a significant conflation effort and is planned for future versions. Another issue is the presence of 1-Many ECRText–GeoNames pairs. In Figure 3, we see example duplications of both ECRText and GeoNames resulting in 1-Many relationships for “Herring River.” With perfect text matches and nearby but differing spatial locations, there is some confirmatory value here, but it is difficult to resolve how to handle these.

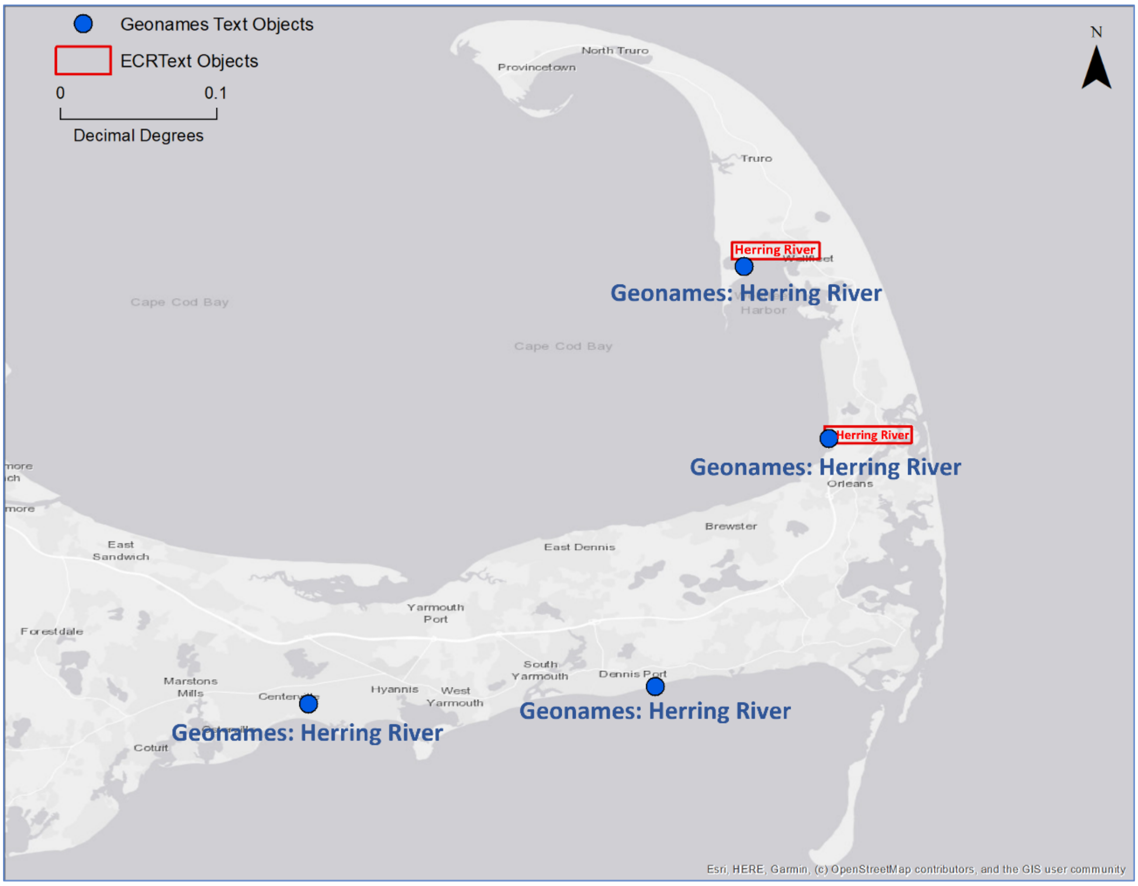

Figure 3: Use of the phrase "Herring River" in GeoNames and ECRText in Cape Cod, Massachusetts.

Conclusion

We show that DNC ECRText data can play a significant validation role for GeoNames data as well as adding new aliases to existing names and adding entirely new locations. We developed a workflow and applied it to compare DNC to GeoNames in the DNC 17 region for classifying ECRText as either confirming or providing aliases to existing named place locations or supplying potential new locations. The implication for practitioners is that ECRText labels are a compatible and complimentary dataset, suitable for enhancing GeoNames workflows operating near or along the East Coast. Future work will address duplication of ECRText within DNC libraries, duplication of geographic names within GeoNames, and the 1-Many ECRText–GeoNames relationships that arise during matching.

Acknowledgements

This manuscript has been authored by UT-Battelle LLC under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan.

References

GeoNames website. https://www.geonames.org/. Accessed June 15, 2022.

“About GeoNames.” https://www.geonames.org/about.html. Accessed June 15, 2022.

International Hydrographic Organization. 2000. IHO Transfer Standard for Digital Hydrographic Data. Monaco: International Hydrographic Bureau. https://iho.int/uploads/user/pubs/standards/s-57/31Main.pdf.

International Hydrographic Organization. 2018. S-100 - Universal Hydrographic Data Model, Edition 4.0.0. Monaco: International Hydrographic Organization. https://iho.int/uploads/user/pubs/standards/s-100/S-100_Ed%204.0.0_Clean_17122018.pdf.

National Geospatial-Intelligence Agency. Digital Nautical Chart. https://dnc.nga.mil/dncp/home.php. Accessed June 15th, 2022

National Oceanic and Atmospheric Administration. Nautical Cartography: The Making of a NOAA Nautical Chart. https://nauticalcharts.noaa.gov/learn/nautical-cartography.html Accessed June 15th, 2022

Kim, Junchul, Maria Vasardani, and Stephan Winter. 2017. “Similarity matching for integrating spatial information extracted from place descriptions.” International Journal of Geographical Information Science. doi:https://doi.org/10.1080/13658816.2016.1188930.

Šimbera, Jan, Dusan Drbohlav, and Přemysl Štych. 2021. “Geocoding Freeform Placenames: An Example of Deciphering the Czech National Immigration Database.” International Journal of Geo-Information 335–351. doi: https://doi.org/10.3390/ijgi10050335.

Dunn, Jonathan. 2020. “Mapping Languages: The Corpus of Global Language Use.” *Language Resources and Evaluation* vol. 54: 999–1018. doi: https://doi.org/10.1007/s10579-020-09489-2.

NGA. 2019. “U.S. Chart No. 1: Symbols, Abbreviations and Terms used on Paper and Electronic Navigational Charts.” https://nauticalcharts.noaa.gov/publications/docs/us-chart-1/ChartNo1.pdf.