Searching for Historical Documents using Combined Cross-Border Geo-Spatial Datasets

Matthew Ng, Master in Computer Science
University of Dublin, Trinity College, 2021

Supervisor: Owen Conlan

This research aims to combine cross-border geo-spatial datasets in order to allow for geo-spatial querying of historical documents across the entire area in which they are associated with. Historical documents were often recorded during time periods in which national borders greatly differed from those of today. The effect of present day borders is that current geo-spatial datasets are often constrained to single jurisdictions. Datasets which cover larger areas lack the potential granularity and historical information that national datasets may provide.

This project combines geo-spatial data into a merged dataset and stores it on a PostgreSQL database which provides coverage for the entire island of Ireland. Features of a geo-spatial search system such as a radius search, filtering by region and determining proximity of locations to a region are sought to be replicated. Documents are linked to geographic toponyms in the merged database through a Levenshtein distance comparison with the place names associated with each document.

The features which were intended to be implemented were done so successfully using geo-spatial queries across the merged dataset. Low recall and precision from the system, however, was caused by the noisy textual nature of the document set as place names were written using non-standardised spellings and were difficult to disambiguate.

A set of guidelines are established for considerations that may be made for merging geo-spatial datasets in the context of using them with historical documents. Through using these guidelines or using the dataset created in this research a basis may be established for constructing a viable geo-spatial search functionality for historical documents if they may be more accurately linked to geographic toponyms.