

Places Matter: Automated Processes for the Analysis and Visualisation of Geospatial Information in Qualitative Data

Martina Tenzer (University of York, Department of Archaeology)

Places matter. Place is a geospatial entity – a physical expression, an environment – but also the repository of collective and individual perceptions, histories, life stories and experiences. It is the space where life takes place, traditions and beliefs find an expression to form communities and identities. Furthermore, place affords qualities for people, and in turn, people shape the landscapes they occupy. People and places are intimately intertwined. Local knowledge and expertise are vital to understanding the particularities of this coexistence and connection and for the planning and decision-making process. Sustainable change and development need public acceptance and civic trust – inclusivity, transparency, and cooperation.

Collecting, analysing, and processing qualitative data, such as free text entries in surveys, social media data, or interview transcripts, is inherently challenging. Assessing and coding qualitative data done manually is a labour-intensive and time-consuming process. Furthermore, identifying and mapping locations mentioned in free, unstructured text is inherently difficult due to the fuzzy and imprecise form of geospatial information provided by such data. Engaging communities and individuals is another challenging factor. Using digital mapping tools for public engagement, for example, Participatory GIS or surveys, including the task of locating places or defining polygons on a map, are still a deterrent for public engagement despite the widespread use of computers, and mapping services, such as Google Maps or Satnavs in daily life.

This paper presents preliminary results of an AHRC/UKRI funded PhD project at the University of York. The project aims to develop automated processes in qualitative data analysis focusing on identifying, locating, and visualising geospatial information in unstructured data. One of the project objectives is to use open-source software, such as QGIS, and R and Python code, to accommodate budget, time and personnel constraints of heritage organisations and local authorities. A sample data set of tweets, interviews and survey data of residents of Sheffield and the Peak District National Park provides the baseline for the methodology of this project.

Current location detection in unstructured text using Natural Entity Recognition (NER) in the Natural Language Toolkit (NLTK) or OpenNLP (Natural Language Processing) is strongly dependent on the data set the model is trained on and the model itself. These data sets are often not sufficiently fine-grained to provide reliable location extraction. We developed a method to extract location data from freely available sources, such as the Ordnance Survey and Historic England. The existing databases were filtered to the study areas and merged in QGIS. The resulting data set provides a gazetteer of over 3,500 points of interest with coordinates. In the next step, this gazetteer is used to detect locations in qualitative data in an automated, time-efficient process in conjunction with an algorithm implemented in Python, which compares tweet content, survey data or interviews with the gazetteer created in the first step and returns a list of matching location entities. While the algorithm handles some of the word ambiguities, for example, common abbreviations of locations or hashtag one-word-style forms of compound location names, it misses specific forms of name alterations, such as misspelt words. Nevertheless, even a simple matching algorithm proves to be accurate to 88%, if used in conjunction with this detailed gazetteer. The standalone application based on this data set provides a method for the detection of predefined locations to create, for example, hotspot maps of visitor impact, highlighting areas of increased demand for nature protection and landscape

management suited to tourism or changing perspectives on specific parts of the landscape. We present applications of the method: visualising a shift in visitor behaviour in the Peak District National Park during and after the Covid restrictions and a hot spot map of resident place attachment based on the automated evaluation of Twitter data and resident surveys.

Further development of the method will include training specialised machine learning models based on this gazetteer of area-specific location entities, to increase the flexibility and ability to identify location ambiguities and test the application of specific heritage and tourism-focused training data sets of similar landscapes, such as other National Parks in the UK. Automated location detection in unstructured qualitative data may offer a time and personnel efficient solution for adding real-time trends and issue awareness into future sustainable landscape development and planning processes.