

Open Source and Open Environment to Enrich Digital Gazetteers and Facilitating Georeferencing Processes

A position Paper to the Workshop on Digital Gazetteer Research and Practice by May Yuan,
University of Oklahoma

Place names (or toponyms) are highly variant and dynamic. Multiple places may have the same name across different administrative areas at a higher order. For example, Miami in Florida is a metropolitan city, but Miami in Oklahoma is a rural town. While both places share the same spelling, their pronunciation is quite distinct: m-ai-a-mi vs. m-i-a-m-ai. Furthermore, a place may have multiple names and local variants, especially for place names that are translated from one language to another. The English spelling for China's capital city can be Peking (an earlier version) or Beijing (the current official spelling). New communities and streets are developed with new names. Existing communities and streets may experience name changes over time.

In addition to city names, names for geographic features (such as mountains and rivers) can change, and as geographic features evolve, locations and geometries associated with these names will change accordingly. Volcanic eruptions or landslides can quickly alter the correspondent morphological and geometrical (shape and spatial extent) associated with these geographic features. New toponyms are given to new geographic features developed by natural or man-made processes (e.g. lagoons, retention ponds). Besides names, relationships among places and/or geographic features can change as well. These relationships may be containmentship, intersection, distance, and other non-spatial or spatial cases. The highly variant and dynamic nature of place names makes it challenging to build a comprehensive digital gazetteer of the world by any one or group of organizations. Hence, the position paper promotes the use of open source information in an open environment to enrich digital gazetteers that take the advantages of rich information from different places, and broad-based local knowledge on the World Wide Web.

Open source information can be on-line or off-line. Off-line open sources are documents or records open for inquires and browsing, such as unclassified government documents, newspaper, books, and other academic literature, etc. While there may be charges to access these off-line open source materials, the information is in general available for the public. The growth of cyberinfrastructure democratizes further publications and dissemination of facts, information, and knowledge. Quality of internet posting, however, varies, but a range of mechanisms has been used to build credibility and reliability. E-commerce and wikipedia are two examples of great success. E-commerce takes user feedbacks to build a reputation for sellers, buyers, or products. Wikipedia, on the other, provides a collaboratory open environment in the cyberspace to build the most comprehensive encyclopedia with the broadest and most diverse author communities. Both e-commerce (reputation established by peer feedbacks) and wikipedia (broad-based authorship) models offer new thoughts to the use of open source and open environment approaches to enriching digital gazetteers. A board-based authorship allows extensive and intensive incorporation of local knowledge that is critical to address the variant and dynamic nature of place names, while peer feedback mechanisms provide a measure of credibility to the authorship and local knowledge. Moreover, a board-based authorship promotes the opportunity to supply historical and geographic contexts to individual place names, most of which have historical, cultural, geographical, or social traces.

Attempts to a broad-based authorship and peer-feedback mechanisms challenge modeling of gazetteer data and meta-data. The traditional alphabet ordering approach cannot effectively handle frequent updates and added information, such as authorship, credibility, spatial footprints (including vector, raster, and imagery data) and versioning, place name time (the time when the place is in use), transaction time (the time when the place name is entered to the gazetteer), and the temporal lineage of place names. Design of such a gazetteer information system shall consider a database with three domains of semantics (e.g. place names, authors, evidence, and contexts), time (e.g. valid time of place names, transaction time of data entries), and space (e.g. spatial footprints, spatial relationships). A semantic object of a place name may be linked to multiple temporal objects of valid time or transaction time, and then linked to multiple spatial objects of footprints and relationships. If a place name corresponds to more than one location, then the place name will be linked to multiple spatial footprints. Similarly, when a place expands (through urban sprawl, for example), the semantic object of a place name will be linked to multiple temporal objects of valid time and then to multiple spatial objects to represent transitions in spatial extents over time. On the other hand, a spatial object may be linked to multiple temporal objects and then to multiple semantic objects to represent a location may have multiple place names over time.

When implemented, advanced search engines are possible to extend place name queries from “where is place name X” and “what is at the place Y” to “How many places have the place name X” and “How has the place changed its name over the years and what is the context for the name changes?” Searches initiated from the spatial domain seek all place names that have been used for a location at the best knowledge of the system. Searches initiated from the semantic domain inquire all locations where a place name has been used to reference these locations. When temporal objects are referenced, searches can be extended to transitions of place names at a location, and an increasing or a decreasing use of a place name over space and time. Subsequent analysis can be done to examine historical and geographic implications for the stability of a place name. Are there regions experiencing common place name changes? Did certain place names become popular in temporal periods or geographic regions? Are certain place names particularly transitory? Are there certain place names commonly embedded special religious, ethnological, social, or functional meanings?

Furthermore, the enriched gazetteer will enhance georeferencing. The gazetteer-based geoferecing process is no longer merely a match between place names and geospatial footprints. Temporal references will indicate how place names evolve at a location or for a geographic feature as well as added historical and geographic contexts to facilitate a better understanding of a place than just its name and geospatial footprint. In addition, the added temporal lineages and geographic context can be used to improve the accuracy of georeferencing, especially when a place name is used by multiple locations or a location has more than one place name. Temporal lineages and historical/geographic contexts help narrow in ambiguity, relate the place to its former or later place names, and associate the place to other place names in its surroundings.

In sum, open source and open environment can offer great promises to expand the current approach to develop digital gazetteers. The expansion can address the challenges of variant and dynamic nature of gazetteers. Advanced place name searches and context analysis of place names can greatly improve the functionality and usefulness of digital gazetteers to facilitate our understanding of places over the world. As the term “place” emphasizes the geographic context of a location, gazetteers built with a broad authorship and local knowledge address directly the

essence of semantic, temporal, spatial components of a place and help us to understand places with historical, geographic, and social contexts.