

## **Gazetteers and Geographical Information Retrieval**

**Chris Jones**

School of Computer Science  
Cardiff University, UK

Gazetteers are coming to play an increasingly important role in geographical information retrieval on the web. They enable users of transport timetables, routefinders, yellow pages, web mapping services and geographical web search engines to employ place names when specifying the geographical context of their requirements. The use of gazetteers in this role has also served to highlight some of their limitations with regard to the needs of the user. In practice, many queries that specify place names fail. One of the prime reasons for this is that the user may employ an informal or vernacular place name that is in common use but which is not recorded in the available gazetteers. In the UK, examples of such names are the “Midlands” the “Chilterns” and the “Wye Valley”. The reason the name would not be recorded is that gazetteers tend to reflect an administrative view of the world with an emphasis upon places that have precise boundaries. Some gazetteers do record the names of topographic features such as mountains and valleys, but they are not usually accompanied by data that record an estimate of the spatial extent of the features. The existing gazetteers may also fail to recognise a name because they lack the required level of detail or geographical extent or simply because they are out of date.

There is a need therefore for richer gazetteers that reflect common knowledge of place names. Because of the high rate at which place names change or are introduced there is also a need to develop a system of interoperable web gazetteer services that reflect local and regional knowledge of places throughout the world.

For the purposes of geographical information retrieval it is possible to envisage an ideal situation in which there is a system of multilingual gazetteer services in which the content conforms to agreed methods for specifying: preferred and alternative names; the timeframe for use of names; an ontology of place categories; rich information on spatial context including geo-political and topographic hierarchies, coordinates in well defined reference systems, spatial relations to adjacent places, and spatial footprints at different levels of generalisation, with information on the nature of boundaries (precise / vague).

### **Vernacular names**

The issue of incorporating vernacular names into gazetteers raises several challenges with regard to the source of the knowledge and methods for modelling and representing it. Where does knowledge of the names come from? There is a great deal of personal knowledge of place names that it is possible to envisage eliciting via some form of mass questionnaire conducted perhaps on the web. There is a considerable body of vernacular place name knowledge within textual documents. Many such documents are to be found on the web and web mining or web harvesting is therefore another route to knowledge acquisition.

Preliminary work on web mining for place name knowledge (by Purves, Clough and others) has demonstrated that simple web queries that include target vernacular place names result in the retrieval of web documents that refer to places that are co-located with the vernacular place, often being inside it. By performing a statistical analysis of the frequency of co-occurrence of place names with the target name it is possible to identify fuzzy regions of space that may approximate the extent of the place. Where the imprecise place contains few other places then it may be appropriate to attempt to identify other co-located topographic features (rivers, mountains, lakes, valleys, forests etc) that may be mentioned in association with the target place.

Several geometric modelling methods have been employed to represent the extent of imprecise places, in particular based upon the use of Voronoi diagrams, Delaunay triangulations (Arampatzis et al) and surface density functions. The latter surface modelling methods are notable for being able to represent the uncertainty of boundaries modelled by the frequency of occurrence of co-located places. Arbitrary precise boundaries can be generated from the surface models by choosing a threshold value of the surface, but it is a challenge to determine what value the appropriate threshold should take.

Building a system of gazetteers with extensive and rich geographical coverage can be expected to require a considerable degree of data integration in order to exploit the variety of data sources available. In addition to methods such as those referred to above, there are more conventional sources of place name knowledge within digital map products, their associated name lists or gazetteers and within geographical thesauri. These sources differ from each other with regard to the coordinate systems, accuracy and precision of geo-referencing, the form of the geometric footprint (points, polygons minimum bounding boxes for example), the nature of administrative hierarchies, the classification systems employed to describe the topographic type of the place, and language variation, along with inconsistencies in spelling and of naming the same places. At present when attempting to merge sources such as the Getty TGN with local gazetteer and topographic map-based names, major problems arise due to differences of the sort listed and there is a need to develop robust methods for integration. It may be useful in the short term to create some benchmark datasets containing expert-asserted equivalences between different representations of places in order to assist in evaluating automated methods for place matching and data integration.

## References

Arampatzis, M. van Kreveld., I. Reinbacher, C.B. Jones, S. Vaid, P.D. Clough, H. Joho, and M. Sanderson, Web-based delineation of imprecise regions, *Computers, Environment and Urban Systems (CEUS)*, Volume 30(4), pp. 436-459.

Purves, R., Clough, P. and Joho, H. (2005), Identifying imprecise regions for geographic information retrieval using the web, *In Proceedings of GIS RESEARCH UK 13th Annual Conference*, Glasgow, UK, pp. 313-318.