

A Mediator-Based Method for Iterative Bottom-Up Gazetteer Generation
[Extended Abstract]

Daniel W. Goldberg
GIS Research Laboratory
University of Southern California

A gazetteer, in the classical sense, provides the foundation for numerous critical services in a variety of academic research fields. Most commonly, a gazetteer's fundamental purpose is to provide the "indirect georeferencing" capabilities that are of utmost importance for applications seeking to use a geographic term for a spatial query (e.g. spatial access to material in a digital library) [1]. Important as this task may be, it represents just one capability for which a gazetteer can be exploited, yet it has led to paradigm of gazetteer generation which produces end results of limited scope usefulness.

This paradigm, which can be termed the "top-down approach", stems directly from the reliance on Hill's "satisficing condition" [1] in gazetteer creation, whereby one must make tradeoffs between the amount of energy expended and the quality of results that are achieved; typically choosing only the most important features for the task at hand (as there are limits to the amount of available time and energy). With this condition, therefore, not all features can be included, and those which are, only up to a threshold of acceptable accuracy. Gazetteer generation efforts using this paradigm, while extremely useful resources for the tasks for which they were designed, are characterized by incompleteness, inaccuracy, and low resolution, partly due to the data sources and methods that are used to create the geographic features themselves. However, more substantively detrimental to the resulting gazetteer, the "top down" paradigm usually proceeds by focusing on a single axis of the gazetteer (i.e. the name, type, or footprint) for which there is an already existing list (typically the name), trying to determine the other two components as generation algorithm proceeds through the list.

In contrast, the research presented in this paper defines a new gazetteer generation paradigm, termed the "bottom-up approach", which focuses foremost on completeness and accuracy, and is not bound by the "satisficing condition". This paradigm enables a conceptual vision of the gazetteer far beyond its most common usage of translating named places into geographical footprints, instead enabling the gazetteer to become an accurate, complete, and authoritative spatial data source in its own right. The key to this approach is releasing the constraint of explicitly defining what constitutes a valid geographic feature suitable for inclusion into a gazetteer, rather, allowing each individual creator of a gazetteer to determine what is suitable and what is not. This is a contentious issue among researchers in the field [2], but if we allow ourselves to make this non-traditional leap, we will enable the gazetteer to become an even more valuable resource than it presently is to a far greater audience.

To facilitate this "bottom-up approach" to gazetteer generation, this paper explores the possibility of employing and extending existing geospatial information mediator architectures [3] to define a generalized framework of the gazetteer creation process, focusing primarily on the ability to easily define and include data sources and operations to be incorporated into the

gazetteer generation process. The primary benefit of such a focus is that one can incrementally improve their gazetteer as new or more accurate data sources become available, algorithms are discovered, and geospatial operations are invented. This framework enables the user to define their own notion of geographic features, and allows non-traditional data sources and spatial operations to be included into the process. Further, because the system is built upon existing geospatial information mediator technology, it is capable of leveraging appropriate data source and operation selection based on availability and suitability of data, as well as choosing the most complete and accurate data sources and operations to use for a particular spatial area (footprint), feature type, or feature name, all of which are determined automatically through logic inherent to the mediator.

To test the feasibility of such an approach, this paper details a prototype implementation of a mediator-based gazetteer generation system for a small section of downtown Los Angeles, California. In this particular instance, the intent is to generate a highly complete (both spatially, i.e. footprints, and in terms of features present) and accurate gazetteer for an urbanized area which could be used in an emergency management situation where the primary query would be “given this polygon area of interest, what geographic features are present, and for each, what is its name, footprint, and type”? This non-typical gazetteer usage represents an ideal case of what a gazetteer generated through a “bottom-up” approach is capable of, which would not have been possible given a traditionally created “top-down” gazetteer, due to lack of spatial or composite feature completeness.

Through an experiment comparing the results of different gazetteers created by employing the same geospatial mediator-based method with different data sources and operations, the results achieved demonstrate that the iterative nature of gradually improving the contents of the gazetteer through the automatic incorporation of higher quality data sources and operations dramatically improve the overall quality of a gazetteer.

[1] Hill, L.L.: Core elements of digital gazetteers: Placenames, categories, and footprints. In Borbinha, J.L., Baker, T., eds.: ECDL '00: Research and Advanced Technology for Digital Libraries, 4th European Conference. Volume 1923 of Lecture Notes in Computer Science., London, UK, Springer (2000) 280–290

[2] Agarwal, P.: Contested nature of place: Knowledge mapping for resolving ontological distinctions between geographical concepts. In Egenhofer, M.J., Freksa, C., Miller, H.J., eds.: GIScience 2004. Volume 3234 of Lecture Notes in Computer Science., Berlin, Springer-Verlag (2004) 1–21

[3] Thakkar, S., Ambite, J.L., Knoblock, C.A.: Composing, optimizing, and executing plans for bioinformatics web services. VLDB Journal 14(3) (2005) 330–353