

Statement of Interest

The Digital Gazetteer Research and Practice workshop interests me for two reasons: digital gazetteers afford the opportunity to expand both the nature and the extent of NGA's support to our customers; in addition, work on gazetteer questions affords an outstanding opportunity to deepen our collective understanding of the nature of and limits to formal descriptions of geospatial data and of commonality between expressions of geospatial knowledge and natural language. The questions posed in the workshop description illustrate the potential for addressing both "applied" and "fundamental" questions within a single, focused body of research.

Components of gazetteer services

The topics identified as focus areas for the workshop are likely to reveal a host of related questions, starting with "what is gazetteer data and, in a digital world, how does it differ from 'regular' geospatial data?" Does the difference lie in the data, in the assumptions that we make about it, or in the kinds of services offered? Which of the following would be gazetteer data: township names; zip codes; telephone dialing codes; designated service areas for commercial services; areas of responsibility for military or non-military organizations.

It might be useful to discuss the expanded utility of gazetteer data provided in digital form, along with scenarios for using new capabilities. Not only will digital services render it faster and easier to perform searches that theoretically might be done with hardcopy gazetteers, new capabilities might support entirely new services and expand the community of users for gazetteers. An extreme case might be to compare an excerpt of spoken language to a list of place names whose various pronunciations are indexed.

Such indexing might require capabilities well beyond today's soundex representations, e.g., application of phonological rules for multiple dialects, in addition to grammar and morphological rules. More straightforward uses might be to find spoken and orthographic variants of a place name (within a designated dialect or sub-dialect or across languages), to find alternative referents for the name and to find (or recognize) alternative names for the "same" place.

Georeferencing as a process

Many of the relationships between bits of linguistic data, and bits of geospatial data, are implicit and context-dependent; a better understanding of such relationships and how to exploit them is essential. Explicit identification of "all" relationships will fail for two reasons: first, the cost of capture and exploitation would be prohibitive; more importantly, we cannot enumerate all of the relationships of interest for a given bit of data.

Traditionally, we have looked for general rules and relationships in grammars and in sets of logical operators that can be combined in useful ways. Linguists impose an "adequacy requirement" on grammars that we do not use in computer science: a grammar should

generate all good constructions in the target language, and it should unfailingly identify faulty constructions as such. What is the adequacy criterion for our work?

Another line of inquiry might be methods and practices for populating and verifying gazetteer data. Are there common parameters we can use to set priorities for tapping into already-compiled (although perhaps not yet digitized) data or for organizing the work? Can we define methods for using the wide range of data becoming accessible to verify, validate, correct and organize place name information? Are there general practices that might be reliable, and can we combine atomic operations into tools to facilitate these practices? Are there ways to tap into user behaviors to find and resolve anomalies in the data? Can we define practical figures of merit for characterizing the quality of data in place name repositories? How do they use or differ from characterizations of quality in “traditional” geospatial data bases?

An ancillary question for anthropologists or geographers might be “how could we use these large, diverse repositories to enhance our understanding of how humans organize place?” In addition to identifying relationships that all cultures capture, are there “potential relationships” that are never used or that always require indirect reference? What is the significance of either finding?

Interoperable gazetteer services

Descriptions of universal underlying topologies for systems more straightforward than place names (e.g., kinship systems) have long eluded anthropologists and linguists. Thus, it is likely that gazetteers will have to be designed to support multiple organizing principles, including mechanisms for handling data that doesn’t fit the schema in use. It might be useful to focus considerable attention on powerful, flexible tools for integrating data from multiple sources. It might also be useful to consider underlying principles for segregating and relating data that we might be able to use repeatedly.

Finally, it would be useful to organize the many classes of questions that this workshop will touch on to capture those where likely answers may lie in other disciplines, those that are dependent on one another, those that must be answered (or assumed), consciously or unconsciously, as gazetteers are built, maintained and used with today’s tools, and research questions that are likely to enable great leaps of progress.

Beth H. Driver