

Expression of Interest in Workshop on Digital Gazetteer Research and Practice

Susan Stone, University of California, Berkeley

My interest in gazetteers is a result of my longtime interest in two different areas of information science and digital libraries research. Since studying for my master's degree in Information Science at Berkeley in 1988 and throughout my work at Berkeley in the Information Server Project in the early 1990s and later the Museum Informatics Project, I have been interested in information system interoperability. My second interest is in language issues and multilingual data. This interest stems from my bachelor's degree in Chinese and my first master's degree in an interdisciplinary program involving sociolinguistics and education, where I specialized in writing systems and scripts.

In the context of the gazetteer workshop, I am most interested in issues involving multilingual and international gazetteer content. In the area of interoperability, I am interested in distributed interoperating gazetteers, gazetteer search and retrieval protocols, and integration of gazetteer data from multiple sources. In my professional role, I am interested in incorporating gazetteer lookup into Berkeley campus collection databases and other information systems. I am also interested in looking at gazetteer development in the context of new technologies and architectures for collaboration.

I first learned about spatial indexing in about 1992, when Linda Hill spoke at a conference of the American Society for Information Science and published a paper in the conference volume, which I edited with Michael Buckland. When I began working the Museum Informatics Project at Berkeley soon after, we used thesauri, such as the Getty Thesaurus of Geographic Names, for standardizing locations in databases of museum collections. We also began to collaborate with the campus GIS center with an eye toward mapping specimen data and other data in campus collections. It was as a participant, representing the Museum Informatics Project, in ECAI, the Electronic Cultural Atlas Initiative, where I actually met Linda Hill and became involved with gazetteers. I participated in ECAI's NSF-funded gazetteer project, which sought to extend the ADL standard to accommodate historical, cultural, and multilingual data. Later, I was part of a research group at Berkeley with Ruth Mostern (then of ECAI, now of UC Merced) and a graduate School of Information Management and Systems student, Melanie Feinberg, to model time period data to complement location data in gazetteers.

In September 2001 through August 2002, NSF funded the project titled "A Multilingual Gazetteer System for Integrating Spatial and Cultural Resources," in which ECAI worked with Academia Sinica of Taiwan to extend existing gazetteers standards to accommodate the cultural and historical resources that ECAI and its collaborators have been developing. For our prototype, we worked on a gazetteer of Taiwan. We worked with location data collected by Academia Sinica field workers, along with the data for Taiwan in the ADL gazetteer and Taiwan data from the Getty Thesaurus of Geographic Names. Data were in Chinese and English, and I developed a testbed in which we were able to search and deliver records from the combined data using the ADL protocol and code from UC Santa Barbara. Part of our work in this project involved investigating a suitable feature type thesaurus for the types of locations researchers in the humanities were most interested in, one that could

be translated into multiple languages (initially, Chinese). Although we did not come to a consensus on how to create a thesaurus that would translate readily into multiple languages, given the different semantic ranges of terms in different languages, one suggestion was to develop a multilingual thesaurus of concepts that could be used to describe locations at a very general level, while also using more specialized thesauri for classification that might not be available or easily translated into multiple languages. Since work on cross-language searching and semantic ontologies has progressed since our NSF project, I believe the issue of multilingual feature types should be revisited in the context of distributed, interoperable gazetteers.

A special focus of the ECAI project was the instability of toponyms over time and the interdependence of time and space in gazetteers. In that regard ECAI came up with suggestions for the ADL gazetteer protocol that involved incorporating time elements into all components of a gazetteer entry, adding more source fields for data acquired from multiple sources, and adding fields for uncertainty—regarding names, feature types, as well as locations. The uncertainty fields that were suggested for researchers working with incomplete descriptions in historical texts proved to be of use for the biological researchers translating location descriptions into spatial coordinates, such as the team now working on the BioGeomancer project, which seeks to create tools for georeferencing and mapping biodiversity data.

Over the past year or so I have been an observer of the work of the BioGeomancer project and have been in contact with some members of the team on language issues. When that project is completed, I would like to see the work they have been doing and the software they develop applied more generally to the groups in the humanities and the collections that I work with in Berkeley's Museum Informatics Project.

In sum, I have been interested in working with gazetteers for a long time and I would like to participate in an effort to create a new research agenda and to put into practice at the University of California the results of current and ongoing research and development efforts.