Workshop Report
Digital Gazetteers: Integration into Distributed Digital Library Services

ACM-IEEE Joint Conference on Digital Libraries
Portland, Oregon
July 18, 2002

By Linda Hill
Alexandria Digital Library Project
University of California, Santa Barbara,
lhill@alexandria.ucsb.edu

A workshop on **Digital Gazetteers: Integration into Distributed Digital Library Services** was held on July 18, 2002 in conjunction with the ACM-IEEE Joint Conference on Digital Libraries in Portland, Oregon. This workshop was sponsored by the Networked Knowledge Organization Systems/Services (NKOS) group as the 5th in their workshop series: see http://nkos.slis.kent.edu/. The 38 participants represented a worldwide community of researchers, implementers, and librarians for whom gazetteers are important components of information systems. These included those from cultural history, biology and environmental data collections, library cooperative projects, digital library research, information service companies, and state and national projects related to gazetteers. Participants came from Germany, Taiwan, Norway, Canada, and the United Kingdom as well as the United States. **Presentation slides from the workshop are available on the NKOS website through the link to the workshop**.

Linda Hill, University of California – Santa Barbara (UCSB), presented the basic elements of a content standard for gazetteer data, pointing out that gazetteers are a type of reference tool along with thesauri, dictionaries, taxonomies, and so forth that are components of digital libraries, valuable for their role in establishing meaning and navigating collections. Details of representing placenames, geographic location (in longitude and latitude coordinates), and type were discussed and issues of fuzziness of place and time and the need for community-based support for gazetteer content standards and typing schemes. Different structures for gazetteer data were compared: thesaurus structure vs. metadata-like structure and the ISO specification vs. the Alexandria Digital Library (ADL) model. The issue of the ranking of places in terms of their "importance" was raised.

Jim Frew, UCSB, presented an overview of the Textural-Geospatial Integration (TGI) Project, part of the NSF's National Science Digital Library initiative. The goals of this project are (1) geospatially-augmented search and (2) geospatially indexing documents. Text documents (includes query statements, metadata, full-text documents, etc.) are the input for TGI. Geographic parsing, thesaurus and gazetteer lookup, analysis of geographic "facts", and evaluation to select the "best" data are the major components of the TGI process. The goal is to have a service to which you can submit a document and

get in return geospatial locations and alternative placenames that are ranked as best, also-rans, and alternatives.

Ruth Mostern, Berkeley, presented the work of the Electronic Cultural Atlas Initiative (ECAI) toward developing digital gazetteer standards for history and culture. They are not focusing on building a gazetteer, but on adapting and developing a content standard, feature type thesaurus, and best practice guidelines for their global, multilingual community of historians and humanities scholars. Language, cultural perspectives, uncertain information about historical places and their names and locations, scholarly documentation, and changes through time are all challenges for developing community standards in this domain.

Greg Janée, UCSB, presented work on the development of gazetteer and thesaurus protocols to support search and retrieval over distributed resources. The gazetteer protocol is also designed to handle the submission of new and modified data to a gazetteer. These protocols are lightweight and stateless and operate over HTTP to encourage implementation. Queries and reports are specified in XML and a small number of basic services are supported. The two protocols are complementary. The ADL model of gazetteer data is used and the model for thesaurus data is based on the NISO Z39.19 standard. The issue of standard relationship types between gazetteer entries was raised: should a small set of basic relationships be adopted? should they reside in the server or the client? should the protocol allow multiple relationship types?

Jens Fitzke, University of Bonn, presented his gazetteer protocol work. It is structured within the framework of the Open GIS Consortium (OGC), which has close liaison with the ISO TC 211 standards work supporting georeferenced data. The OGC Web Gazetteer Service is built on the OGC Web Feature Service and adopts the model of gazetteer specified in the ISO TC 211 DIS 19112 standard for spatial referencing by geographic identifiers. Unresolved issues include handling feature changes over time. The issue of how to harmonize the two gazetteer service approaches was raised.

Dagobert Soergel presented a report on current NKOS activities sent by Gail Hodge. Of note is that NISO is planning to revamp the wording of their Z39.19 thesaurus guidelines to be more accessible to non-lexicographers and to reissue the standard without extensive changes. Discussion questioned the wisdom of this move since discussions have already been held to identify the important changes needed for the standard.

Ya-ning Chen (Arthur Chen), Academia Sinica, presented work toward a digital gazetteer service in the context of Chinese culture. He compared the support for gazetteer information in three standards representing the approach of a digital library, a museum, and the library world: the ADL Gazetteer Content Standard (GCS), the Thesaurus of Geographic Names (Getty), and USMARC standard. They have translated much of the GCS into Chinese and have added links into a GIS system and to the map layers that contain the place, perhaps at different scales of resolution. Representation of change in names and in location through time is an important factor. Results of this work include an online Taiwan gazetteer, online Chinese Civilization in Time and Space, an XML testbed

for the GCS, a Chinese version of the GCS, and a set of feature types for Taiwan. Future work includes development of a hybrid approach to the use of GCS, TGN, and USMARC gazetteer data; gazetteer sharing services for file exchange and distributed retrieval, and the development of relationship attributes for gazetteer entries.

James Reed and Andy Corbett, Edinburgh Data and Information Access (EDINA), presented the UK Geo-Crosswalk project whose aim is to develop gazetteer services for the UK tertiary education and research community. Their approach, based on ADL, is to support geo-parsing and enhanced geospatial searching and provide reference services, metadata creation services, and tools for spatial searching for the academic community and beyond. Technical issues stem from the variety of sources that need to be used from placename and geometry data and include complexity, feature typing, positional accuracy, alternative names, time stamping, and incompleteness of data. Licensing is also an issue. They have developed their own spatial searching software which is being used with an Ingress database.

Humphrey Southall, University of Portsmouth, presented the Great Britain Historical Boundary Project which includes a gazetteer component but is much more than that. This raises issues of the degree to which their gazetteer component can be compliant with the ADL GCS. A fundamental problem with the analysis of social statistics through time is the complexity of changes in administrative boundary and organizational hierarchies. This project seeks to improve the metadata documenting these boundary changes and to develop an integrated and comprehensive gazetteer/place-name authority list for historical administrative units in Great Britain, linking information about the same place from different sources (primarily census data). They are basing their system on an Oracle database and their spatial search datablade and GeoTools open-source visualization toolkit for the client. Complexity and abundance of the census data available is a major issue, as is adopting the social science Data Documentation Initiative (DDI) standard for representing details of survey data. Other issues include anticipating the number of users and their uses for their website; implementing ADL GCS in such a large, complex system with time-variant footprints; developing a set of feature types for their administrative categories; language-specific preferred names; supporting both longitude/latitude and grid spatial geometries; the degree to which the statistical data resides in or out of the gazetteer structure; and explicit relationship types among entries.

David Smith, Tuffs University, presented his work on mining gazetteer data from digital library collections done as part of the Perseus Project. Here one challenge is to recognize placenames in parallel corpora in different languages; e.g., an English translation of an original Greek document. Knowing this mapping is an aid to aligning the text of the two documents. In a fashion similar to the way in which words and their meanings are extracted from texts for dictionaries (a process known as “slipping”), the proposal here is for text to be mined for their placenames and variant forms of the placenames, especially variant language expressions for those placenames, and to populate gazetteers with this knowledge. Interesting statistics on the nature of placenames has come out of their work, showing cultural/locational differences in the frequency of multiple names for the same place versus the frequency of the same name designating different places. Recall and

precision measures on the effectiveness of their methods for identifying placenames in text in different languages show high performance in the range of 0.89 to 0.96 (F measure combining recall and precision measures).

Patrick McGlamery, University of Connecticut, presented his work involving the use of digital cartographic sources to estimate the geographic extents for 718 named populated places in Connecticut, some of which have naturally indistinct boundaries that cannot be obtained from official sources. The process involved starting with the populated place name set from the USGS GNIS dataset for Connecticut, which includes point location data only, and a set of satellite imagery which has been processed to represent Land Use / Land Cover (LULC). The LULC data classify the terrain into classes; two of the classes, "residential" and "rural residential," were used for this analysis. The process resulted in 258 successful matches of GNIS points with single residential polygons. Unsuccessful matches included 193 instances of multiple points per residential area and 267 points with no corresponding residential polygons. Many of the latter places were named "Corners" indicating, perhaps, minor populated places at road intersections. The process is considered to be a satisfactory way to obtain polygon boundaries for gazetteer data, but it must include local authentication. Since there were also residential areas identified for which there was no GNIS populated place entry, additional analysis into if or how they are included in GNIS is necessary.

Rhian Evans, Atlas of Canada, presented the current activities of the Atlas of Canada project and some of the issues they have encountered that are related to their gazetteer services and “find a place” application. Both the Atlas project and the Canadian Geographic Names Service (CGNS) operate under the umbrella of the Canadian Geospatial Data Infrastructure (CGDI). Challenges for the Atlas project include problems with the structure and content of the CGNS and duplicate names for the same features that show up in other sources; relationships between the data in different GIS layers; features with fuzzy boundaries (e.g., bays); and data tagged at different scales. The CGDI gazetteer services attempts to perform “on the fly” linking of placenames to feature data in the Atlas. Multiple sources of data (e.g., census data, environmental data) are linked to a particular feature. Future plans include adding names for features such as health districts and ecozones and adding additional thematic layers to the Atlas.

Cliff Lynch, executive director of the Coalition for Networked Information (CNI), summed up the workshop and led the discussion period. He began by pointing out the tension that gazetteers present between the view that language (placenames) is a nuisance to real geography and the view that named geographic places represent more than geometric inclusion. He cautioned against the desire of some to have gazetteers include encyclopaedic knowledge about places rather than serving their role as reference sources that link or provide access to extensive datasets and publications about those places. There are nasty engineering problems to deal with, including persistent IDs when referencing texts. An important short-term goal should be to deal with the issues of distributed gazetteers, dealing with issues of scooping, provenance, and versioning. He cautioned against combining in one gazetteer service protocol the search and retrieval functions with the submission and update functions; much more agreement among parties

is required for update than for retrieval. There are questions about how much context to carry with gazetteer queries. Experience with Z39.50 should be considered; moving elaborate semantics can be tricky, with implications for processing speed and usefulness. On standards development, he supported the role of ground-up standards processes where applications and implementations make the case for adoption by others. A killer application can make the case for such a standard. He pointed out that simple visualization of documents geographically is a mere party trick; the real value of georeferenced information services lies in finding associated concepts and doing query expansion and assisting in cataloguing and metadata creation. He suggested that the National Science Digital Library initiative offers an opportunity for local gazetteer development driven by the needs of education. Data quality issues that are revealed through geographic access is an intriguing issue.

The discussion period was lively and can't be completely summarized here. Topics included (1) security and privacy issues with gazetteer data; (2) importance of feature type schemes for information description and retrieval; need for customized/localized feature type schemes; interoperability of such schemes; tension between complexity and simplicity in these schemes; (3) announcement about the ESRI gazetteer services available free over the Internet and the distinction between placenames and addresses – geoparsing vs. geocoding; and (4) interest of CNI in gazetteers; increasing interest in structured reference sources.

In addition to the workshop proceedings, Jordan Hastings, UCSB, gave a presentation on duplicate detection challenges for gazetteers at a Workshop Preview session the evening before the workshop. He presented his work on developing a methodology for comparing gazetteer entries to determine if they are about the same place. This is a difficult challenge because no single piece of data about a place is unique: the same names can be linked to different places; a place can have multiple names (sometimes because of different languages, variant spellings, and changes through time); geometries come in different forms (e.g., points and polygons) and at different scales, from different suppliers, and for different time periods. Types can also vary, and have differing hierarchical levels of representation, but have proved to be the most reliable clue for duplicate detection. The methodology uses a weighted combination of measures based on a comparison of types, locations, and names. Other factors that might improve performance is whether the feature can be expected to have a crisp or diffuse boundary; whether the feature's geometry is compact or extended; and whether the feature is tangible or abstract. Map visualization of potential duplicates, with appropriate background layers, aids in human evaluation. The database challenges for this processing are significant, calling for custom data types, multiple (unlimited) attribution, and efficient geospatial processing. A proposed processing cycle is described.