Digital Gazetteers for Reference Mapping Applications
Position Statement, Digital Gazetteer Research & Practice Workshop
Santa Barbara, CA   December, 2006

Howard Veregin, Ph.D.
Director, GIS Operations
Rand McNally & Company

In many ways a digital gazetteer is no different than a GIS feature class. The basic components of the gazetteer – name, type, and location – are sufficient to enable simple reference maps to be made with labeled features and symbology differentiated by feature type. Additional attribution – population, for example – allows for more complex map representations. Adding a capability to handle multiple names and multiple locations allows the map representation to be tailored to particular map specifications or product editorial guidelines. By accommodating parent-child information linking features at different levels of the geographic hierarchy, features can be selected for different types of cartographic treatment or for different naming policies.

Viewing a gazetteer as a support tool for reference map applications exposes some content and design issues that might otherwise be overlooked. This position paper discusses some of these issues with specific reference to mapping requirements for US populated places.

Is This Place Real?

Grovers Mill, NJ, is the fictional landing site of the Martian invasion in Orson Wells' *War of the Worlds* radio broadcast of 1938. It is also a small unincorporated place of about 120 people close to Princeton, with a history that goes back to the 1700s. Like Grovers Mill, most places in this country have an existence that is part real and part imaginary. Sometimes this is due to change over time – places are abandoned, or annexed by other places, or end up at the bottom of a new reservoir. Many gazetteers are littered with such "dead" entities.

Legal status also plays a role in real versus imagined existence. Incorporated places exist unambiguously because they have a legal boundary and governmental functions. Unincorporated places have neither a boundary nor a government, and some are more "real" than others. The US Census Bureau delineates CDPs (Census Designated Places) for significant population clusters that have no legal status. Some are decidedly real, such as Arlington, VA, which has a population of almost 200,000 and a name that is widely-recognized. Other CDPs, however, are for statistical purposes only. They are population clusters but have little or no local meaning or identity. For example it is doubtful that the inhabitants of the CDP called Hickam Housing, HI, have ever heard of the place.

In map production, place validity is closely tied to product specifications and editorial guidelines. For example, on a road map, all places should have some discernible physical presence on the ground that can be used for navigation. A sizable population might not qualify if it is scattered over a large area. A small population cluster might suffice if there is a cluster of houses, a fire station, or a water tower with "Welcome to Waldo, Wisconsin!" painted on its side. Attributes reflecting size or importance, such as population, are needed to identify which places take priority on a small-scale map where space is an issue, and also to tell the map user where she might find a gas station or convenience store. Historic places might be shown as a POI but not as a populated place, and statistical areas like Hickam Housing would not be shown at all.

Places do not just appear and disappear over time, but also undergo transformations. Until its incorporation in 2001, Goleta, CA, was described by the Census Bureau as a CDP with a population of over 55,000. Post-incorporation, its population dropped to around 30,000 because the footprint of the incorporated place is smaller than that of the CDP. The implications of Goleta's incorporation include possible changes to its map symbols (town dot and text size) and index entry (if population is shown). In extreme cases feature selection itself may be affected.

As these examples show, reference map applications require a rich set of feature attribution related to place type, incorporation status, historical status, local recognition, population, and so on. Since this information is expensive and difficult to collect and maintain, many existing digital gazetteers do not provide it. For example, in GNIS the basic place-type categorization includes only populated places and locales, with an optional historic designation. As such much additional research and development must occur before GNIS data can be used to support reference mapping activities.

Duplication and Apparent Duplication

In this country, place name duplication is the norm, not the exception. One reason is the relationship between intersecting administrative units, the most obvious example being municipalities (cities, towns, villages) and MCDs ("towns" and townships). Coupled with this is the existence of various official agencies, like the Census Bureau, which create areas for statistical reporting at various levels of the geographic hierarchy.

As a result, Groton, CT – an incorporated place with a population of about 10,000 – co-exists alongside Groton, CT – an MCD ("town") with a population of about 40,000. The MCD has a larger population because it covers a larger area. While these two entities have the same name (ignoring for now the issue of how the generic term "Town" is handled) they have different legal status, which is important in reference mapping due to the need to ensure that product specifications are being adhered to. If nothing else, given the different populations of these two entities, choosing one over the other would affect the dot and text size on the map.

A similar type of confusion occurs between CDPs and other delineations of unincorporated places. While CDPs are constrained along the boundaries of census enumeration units, there is no reason why this must be so. Landscan and similar population grids allow us to define population clusters more arbitrarily, which can give rise to multiple views of the same population cluster, all potentially valid relative to definitions and criteria.

A more difficult problem is the inverse – places that are truly duplicated in the database, but have not been identified as duplicates because their names are different. It is relatively easy to de-dup Goldengate, IL and Golden Gate, IL. More difficult is Branson West, MO, which is also known as Lakeview. Most places in the US have multiple place names they inherit from common usage or official status, or that have been designated by the US Postal Service or Census Bureau.

The goal of having one entry for each named place oversimplifies reality – or at least the reality of the US populated place landscape. Multiple records for named places is a common occurrence. If de-duping of records is to occur, it needs to be based on more than just name and location, as this will undoubtedly degrade the richness of the data to some degree. For reference mapping purposes it is more desirable to retain multiple records if there is some logical reason for them, such as different legal status, as long as the gazetteer provides the necessary attribution to differentiate these records, and as long as the alternate names problem is properly handled.

Geographic Hierarchies

Explicit information about hierarchical relationships between geographic features is a necessity for map production tasks. If these relationships were not available cartographers would not be able to perform even simple functions like sorting places by state to create a state-by-state index, or performing feature selection to customize symbology or naming conventions in different regions.

In the US, hierarchical relationships can be complicated, and simple spatial point-in-polygon rules often fail to provide expected results. Places can have multiple affiliations, such as Buffalo Grove, IL, which crosses the Cook/Lake county line. The same is true at the MCD level. To correctly associate places with their MCD and county "parents" it is necessary to compare polygonal representations, which can be tricky given the slivers and gaps that result from different spatial representations and levels of generalization.

Even with polygonal representations, errors can occur. In about half of the "township states" (itself a fuzzy set) there is no overlap between municipal and MCD governments. An example is Wisconsin, where municipalities and MCDs are mutually exclusive from a legal point of view. In the other half of the township states, some (but not all) of the municipalities within the state operate within territory that is also served by an MCD government. Some governmental functions are the responsibility of the municipality and some are the responsibility of the MCD. In Illinois, for example, all municipalities are within an MCD except Chicago, Cicero, and those municipalities in counties that have no MCDs. If product specs call for differentiation by governmental authority, it may be necessary to rely on other sources (e.g., the Census of Local Governments) if accurate information is not available in gazetteer format.

Or consider Hawaii, which has no legally-defined cities at all, except Honolulu, which has a single municipal government exercising control over the entire island of Oahu. The Census Bureau recognizes a CDP called Honolulu that is much smaller in size that the city/county of Honolulu. One issue is whether smaller unincorporated places on

the island of Oahu are "part of" Honolulu or not. This would affect how they would be symbolized in a product, or if they were selected for display at all.

Existing digital gazetteers take different approaches to displaying hierarchical information, from text-based explanations (GNIS) to highly structured hierarchies (Getty Thesaurus). Few gazetteers provide hierarchical information past the county level, and information is sometimes obsolete, a reflection of the expense and difficulty of maintaining data to this level of detail.

Conclusion

Reference map production is obviously just one application of gazetteer data, but it does offer some unique perspectives. A fuller understanding of the issues requires an analysis that goes beyond the US and beyond populated places alone, to encompass a broader geographical context. Still there is no question that to provide enhanced utility for detailed reference mapping purposes, digital gazetteers require flexibility, accommodation of local geographic variables, and attribute richness to try to capture the complexity of the real world.