An Area-Based Encoding Scheme for Place Names

Kirk L. Duffin

28 Feb. 2002

Extended Abstract

Place names in genealogical research have a dual purpose. The first is simply to show where an event occurred, which is often interesting in its own right. The second purpose, even more important from a research perspective, is to localize sources of records. Knowing what sources of records are available for a given area or neighboring areas can lead to more information about a particular person or family.

Place names are compact and well suited for genealogy work. However the place name is simply a label, referring to a specific geographic area. This is sufficient for human researchers, who use the label to access mental maps and printed maps.

It is extremely important to note that in transferring genealogical information to the computer, the place name labels are recorded, but the implicit references to geographic area are not. In essence, the label becomes the representation. If effective automated research tools are to be created, the machines must be given the means to understand the underlying geospatial relationships.

This paper examines the weaknesses of traditional place name notation as place representation and presents a method for encoding the location associated with the name. The encoding method is based on triangular quadtrees of the faces of a spherical octahedron representing the globe.

1 Hierarchical Place Names

The most traditional place name style used in genealogical research is the three level hierarchical name, typically a town, county, state combination.

The hierarchical place name is compact and works well for machine encoding. However it has some weaknesses, which include

- Precise names required. Doesn't work well with ambiguity, e.g. "5 mi. north of town".

- Extensibility. Not everything fits well within a three level hierarchy. In the United States, the three levels are typically town, county and state.

With typical American hubris the nationality is assumed. In European research, the middle field is often used for provinces or shires, politically equivalent to states and the last field is used to denote the country.

Many records sources have additional hierarchy levels associated with them, giving rise to the need to record parish and township information. Unfortunately, a typical solution to this problem in computerized records is to eliminate the offending field.

- Comparisons. Unless everyone agrees on the hierarchy levels and spellings used for a particular place, comparisons and consequently searches are difficult. Consider the set `Salt Lake City, Salt Lake, Utah`; `Salt Lake, Salt Lake, Utah`; `Salt Lake` (city or county?)`, Utah, USA`, etc.

  This is further complicated by the fact that even the hierarchical ordering relationship between geographic entities is not absolute. In urban areas, a single city may encompass multiple townships, while in rural areas a township contains several towns.

- Lack of geographic locality. The hierarchical place name system does not lend itself to describing distance relationships. For example, based on names alone, there is no indication that `Nauvoo, Hancock, IL` is merely a few hundred yards distant from `Montrose, Lee, IA`.

2 Geographic Information Systems

One solution to the weaknesses of the hierarchical place name system is to couple the names to a geographic information system (GIS). GIS systems are excellent for analyzing geographic relationships but are not well suited to the needs of genealogy research.

GIS information is usually encoded as sets of high precision coordinate points. This works well for small details but does not lend itself to area relationships. As an example the U.S. government has provided the Geographic Name Information System (GNIS) which gives standardized names, latitudes, and longitudes for hundreds of places in every state. However, representing an entire city as a single point loses considerable information about what actually belongs to the city. Representing a county or a state as a single point is even more ludicrous.

GIS systems also typically organize sets of points into boundary descriptions which can describe areas. The boundary representations are extremely compact for the amount of detail they contain and they are extremely precise in separating one region from another.

The biggest weaknesses of GIS systems are level of detail and ambiguity. In the boundary representation, the amount of detail used to represent one area may be totally inappropriate for another level. Representing the boundary of a city to 10 meter resolution may be appropriate, but representing a state at the same resolution is not. Schemes for simplifying curves exist, but the results tend to be highly subjective.

3 Quadtrees

Another representation of area coverage, used in computer graphics, is that of the *quadtree*. A quadtree starts with a square which surrounds the object of interest. This square represents the root node of the quadtree and is marked as containing part of the object. The square is then subdivided into four equal pieces by connecting the midpoints of opposite sides. These four subsquares are marked according to whether or not they are in the object. These subsquares are the children of the root node. The subsquares are recursively divided as many times as needed to represent the object to a desired level of detail.

The quadtree representation is succinct. Only subsquares containing the boundary of the object need to be subdivided further for more detail. Subsquares completely inside or outside the object do not need to be further subdivided.

Furthermore, the quadtree scales well. The subdivision can be terminated at any level and still result in an accurate approximation for that level of detail. In addition, further refinement to the boundaries can be added without affecting the higher levels of the tree.

4 Octahedral Quadtrees

The square is not the only primitive in two dimensions that can be subdivided self similarly. A triangle can also be subdivided into 4 pieces by connecting the midpoints of the sides.

This is important because of the irregularities in a square grid applied to the surface of the globe. As the poles are approached, the amount of distance covered by a unit of latitude remains roughly constant while the distance covered by a unit of longitude decreases drastically. In contrast the parameterization of a triangular grid remains roughly constant over the surface of the globe.

The geometry used in the spatial encoding scheme presented here is that of an octahedron, projected to the surface of the globe. The triangular faces of the octahedron are subdivided using the quadtree approach described above.

For external representation, the quadtrees are scanned in breadth first order so that the higher nodes in the tree come first in the data stream. This has the benefit that the data stream is incremental. A place representation can be truncated at any point and still remain a valid spatial encoding. Additional data stream information only refines the representation.

5 Operations

This paper will also examine various operations that can be performed on the quadtrees, namely intersection, union, and tests for inclusion.