

On Gazetteers

John R. Frank

MetaCarta Founder and CTO

November 2006

MetaCarta sells commercial geoparsing and geographic information retrieval systems. These products depend on data bundles called Geographic Data Modules (GDMs). Each GDM contains a gazetteer and a set of linguistic statistics that model the natural language usage of geographic references. These linguistic statistics capture nuances of human discourse. The MetaCarta GeoParser uses its GDMs to predict what phrases a human would extract from a document and resolve to particular locations. The MetaCarta Geographic Text Search (GTS) system indexes geographic and textual information from large collections of documents, so users can find everything known about any place.

Anyone can generate new geographic names simply by using them in communication. From the perspective of adding geographic structure to unstructured documents, all geographic references are valid. One has no recourse to enforce particular naming conventions. Authors' choices dictate. Thus, gazetteers are unbounded in scope.

While anyone can create new georefs, intuition suggests that a majority of literate people agree on a special subset of common locations. Plotting the frequencies of occurrence of georefs from a balanced corpus of data partially confirms this expectation. The distribution has a peak of high frequency names, but the distribution is fat-tailed. Most gazetteer efforts start in the peak and grow by expanding toward the tail.

Experience with peak frequency names underscores the value in the diverse swarm of locations beyond the peak. Plenty of structure remains outside the peak. For example, power law models might illuminate the social processes that bring a location into awareness.

Further segmentation is useful. MetaCarta organizes GDMs by language and genre of discourse. Frequency distributions vary across languages. Various genres use location names differently. For example, news articles often uses geographic names as metonyms for state actors. Short designators like "Building 4" or "Lease Block 22" appear frequently in corporate and technical discourse. Style and tone change the meaning of georefs in government reports, grade school texts, geology research, travel guides, and other genres.

This philosophical question deserves attention: What defines a location? Can we rely on a Platonic form for Location? If France, the Statue of Liberty, and the North Pole are to be instances of Location, then its definition must encompass many ideas at once.

Most uses of the word "location" imply a geospatial sense, and yet, geometry alone fails to capture the full meaning ascribed to a location. Simply listing polygon vertices

does not capture the essence of what people mean by locations like Aix-en-Provence or Rome or the Great Wall. One naturally thinks of history, political hierarchy, and other relationships that mankind has with that entity.

Is the oil platform known as "Jack" simply a longitude-latitude bounding box? Oil platforms' most striking features are their drills stretching into the depths and changing orientation during operation. Besides this geometric complexity, when such an entity enters the consciousness of more people, its name takes on new meaning. Just as Cazumel means "vacation" in a somewhat generic sense, the discovery of oil at Jack has taken on significance independent of its literal location – for some people.

In the process of communicating, people add attributes to location entities. As a location gains importance, communities wander away from asserting rigorous gazetteer definitions. Definitions escape. The line between defining and describing fades.

Locations often come into existence as bookkeeping conveniences. As they progress from mere gazetteer entries up the fat-tailed distribution of awareness, locations gain personality. Only natural language can capture the richness of these attributes.

What defines a location? The collective awareness of many people defines a location. Structured gazetteers cannot capture this – at least not without true artificial intelligence.

The emerging field of geographic information retrieval offers a pragmatic way forward. Instead of engineering more and more knowledge into gazetteers, we gain more ground by acknowledging the boundary where gazetteers stop and unstructured GIR takes over.

Gazetteers form a fundamental part of GIR. Articulating useful boundaries between these two types of tools will foster progress in both areas of effort. GIR offers a bridge toward cartography and information presentation, which connects gazetteer efforts to people whose awareness drives the organization of gazetteers.