



Geocoding and Business Intelligence

By Dr. Ramon Barquin and Diogenes Torres

Geocoding and Business Intelligence

By Dr. Ramon Barquin and Diogenes Torres

Have you ever sensed the frustration of a TV detective who finds a scrap of paper with an incomplete address (i.e., missing city or state) that may just allow him to crack a case or prevent a new crime and pursues a number of false leads until finally getting to the right place and nabbing the bad guy? If so, you are just getting a glimpse of the type of gyrations and desperate maneuvers that real investigators in law enforcement or homeland security agencies go through as they try to extract business intelligence from geospatial data tied to crime prevention or the war on terrorism. The FBI, the Drug Enforcement Administration, the Secret Service, the Department of Homeland Security's Immigration and Customs Enforcement component are all frequently in this boat.

But the need to mine business intelligence from geospatial data doesn't stop there. It is difficult to think of a single federal agency that does not at some point require an in-depth utilization of geographic information. The Department of Housing and Urban Development (HUD) often must know exactly where a specific property is located to determine whether it lies in an area that has been designated for special statutory treatment, such as Section 8 rent subsidies, which are defined through city-based "fair market rates." Congressional assistance after major emergencies is usually distributed by naming counties as disaster areas, and the IRS will often provide tax breaks for damages arising from that emergency. The Small Business Administration has to determine whether a specific business is headquartered within the boundaries of a geographical area in order to establish whether they qualify for "HUB Zone" (impoverished location) status and are hence eligible for special treatment. The EPA provides for a special designation of areas where hazardous waste or other types of toxic substance have contaminated the soil or the water supply, entitling residents to certain federal benefits including tax deductions.

But how do we know in any of these cases, whether a specific household, person or business, are eligible for the benefit? First, it has to be determined whether a location falls within the designated geography. That may be easy to determine if an address has a zip code that is clearly within, say, a well established core-based statistical area (CBSA), but if there are several hundred or thousands of addresses that need to be verified for location and to determine whether they are within a specified area, it can get very dicey. This is where the importance of geocoding becomes clear.

Geocoding is the process whereby, knowing an address (street, number, city, state), a set of geographic coordinates are assigned to said address thus marking, within a degree of precision determined by the instrumentation, its specific location. This marker is defined by a pair of numbers that provide its latitude and longitude (lat/long). And given that all geographies are defined by polygons with preset borders, having the lat/long of any address allows us to perform

the geometrical mathematics that let us establish whether that address falls inside or outside the geography of interest.

But geocoding is not easily done. Determining the precise lat/longs of buildings, landmarks, houses, plants or power stations requires that someone go there with a transponder or GPS and capture the specific coordinates. Furthermore, what coordinates are we talking about? A building may have several corners that define its base or several other corners that define its rooftop. A power station may include a complex of several buildings, substations, parking spaces and open areas. And, because it is not likely that anyone will ever actually obtain the coordinates for every single structure in the country, there are established approaches to extrapolate the lat/longs of an address that is toward the middle of a street (center line) when all one has are the coordinates for the first and last house in a block. The dominant company dedicated to capturing geographical information, including latitudes and longitudes in the U.S. is Chicago-based NAVTEQ, and one of its principal competitors is Tele Atlas, a Dutch firm.

Geographical information systems (GISs) have been around for several decades, but with the advent of services such as Google Maps and Google Earth and the proliferation of mobile computing and location aware devices, the need has intensified to develop platforms for producing geospatial business intelligence. And, as mentioned, governments in particular seem to have an insatiable appetite for these applications.

In order to address these needs, agencies are waking to the importance of hosting enterprise-wide geocoding support operations with a variety of functions. A recent request for information (RFI) issued by a federal agency, for example, outlined the following requirements:

- Ad hoc street address matching to include validation, standardization and geocoding.
- Batch processing of street addresses to include validation, standardization, and geocoding.
- Address validation, certification, correction and other interpolative capabilities.
- Methods of determining and matching the location of landmarks, cities, towns and other commonly used geographic references (i.e., “Gazetteer”-level functionality).
- Proximal intelligence or supplemental detail on environs of geocoded results beyond simple latitude and longitude returns (i.e., demographics, political boundaries, census designation, congressional districts, etc.).
- Ability to load and manage custom taxonomies, data dictionaries and GIS data sets.
- Reverse geocoding capabilities to translate incoming coordinate data and dimensionalize that geo-location data to a local street address, nearest populated place, etc.

Because there is no one single commercial off-the-shelf (COTS) tool that will meet all the classical geocoding requirements, most agencies will require a portfolio of several software tools

available in the market that address some of these requirements. Each has a specific range of coverage such as: geocoding, geographic entity extraction from structured and unstructured data, geographic visualization, Gazetteer capability, etc. In order to obtain a robust full-service geocoding platform, agencies are looking into creating what could be called a Location Expert and Geocoding Service (LEGS). In addition to the tools required, a LEGS will surely necessitate a certain amount of custom application development to extract and store information that is not readily available through these tools.

A robust LEGS will consist of several components. The main ones are:

- Data input and output,
- Geocode databases, and
- Geocoding services and architecture.

The principal services provided by a LEGS could be subsumed under the term “address enrichment services” and are:

1. Data quality, address standardization service
2. Geocoding service
3. Geospatial service
4. Mapping service
5. Demographic service

The Department of Housing and Urban Development probably pioneered some of this 20 years ago by creating a Geocode Service Center (GSC) that standardizes and enriches close to 100 million addresses per year. While the HUD GSC functionality is somewhat unsophisticated by today’s standards for a LEGS, other federal agencies are only now stepping up to the need to provide such enterprise-wide support.

Geospatial business intelligence (GBI) is essential for governments in the 21st century, and in order to do GBI well, there has to be a LEGS or its equivalent within every public sector enterprise.

About the Authors

Dr. Ramon Barquin is the President of Barquin International, a consulting firm, since 1994. He specializes in developing information systems strategies, particularly data warehousing, customer relationship management, business intelligence and knowledge management, for public

and private sector enterprises. He has consulted for the U.S. Military, many government agencies and international governments and corporations.

Dr. Barquin is a member of the E-Gov (Electronic Government) Advisory Board, and chair of its knowledge management conference series; member of the Digital Government Institute Advisory Board; and has been the Program Chair for E-Government and Knowledge Management programs at the Brookings Institution. He was also the co-founder and first president of The Data Warehousing Institute, and president of the Computer Ethics Institute. His PhD is from MIT. Dr. Barquin can be reached at rbarquin@barquin.com.

Diogenes Torres is Chief Data Warehousing Architect at Barquin International. A recognized expert in databases, data warehousing architecture and geospatial processing systems, he has taught the data warehousing consultants at Oracle, BearingPoint/KPMG, Deloitte & Touche and PricewaterhouseCoopers. He has been a member of the faculty at The Data Warehousing Institute, where he was a regular from the early conferences. He had long careers with NCR and DEC, and for the last ten years has been working with Barquin International.

He designed the modernization of the HUD Geocoding Service Center and has consulted for suppliers of geocoding software. Other engagements in federal business intelligence projects include the design and implementation of HUD's Multifamily Housing data warehouse, architecting the Advanced Query Facility for Census 2000 and serving as lead architect in several Department of Homeland Security data warehousing efforts.