DESPLIEGUE DE UN SERVICIO NACIONAL DE GEOCODIFICACIÓN. EXPERIENCIA CUBANA.

Carlos José de Armas García, Andrei Abel Cruz Gutiérrez

Resumen. La geocodificación es el proceso de encontrar la ubicación que corresponde a una dirección postal o nombre geográfico, y se realiza comparando los elementos descriptivos de la dirección con aquellos presentes en una base de datos de referencia. Este proceso puede integrarse a múltiples aplicaciones en las que se requiera la posición en el mapa de los datos de entrada. Con esto se pueden realizar entonces análisis espaciales de la información, por ejemplo, establecer la relación entre determinado contaminante y la prevalencia de cierta enfermedad. En este trabajo se presentan los resultados del proceso de implementación y despliegue de un servicio de geocodificación, con cobertura nacional, en el contexto de la Infraestructura de Datos Espaciales para el enfrentamiento del delito de la República de Cuba. Como ejemplo de utilización masiva, se exponen las estadísticas de la aplicación del servicio para la geocodificación del Registro de Electores a nivel nacional.

INTRODUCCIÓN.

Los Sistemas de Información Geográfica (GIS, por sus siglas en inglés) se han convertido, desde finales del siglo pasado, en una de las tecnologías de más amplio impacto en todos los ámbitos de la sociedad moderna, desde el ciudadano hasta los gobiernos, y sobre todo en estos últimos.

La caracterización espacial de los procesos tanto naturales como sociales resulta decisiva en infinidad de áreas de la actividad humana. Así, la implementación de complejos análisis espaciales ha contribuido a identificar las causas asociadas a fenómenos como la aparición de una peligrosa epidemia, la elevada proporción de casos con cierto padecimiento o el incremento del delito en una zona residencial determinada.

Por otra parte, la gestión y optimización de sofisticados sistemas de transporte resultaría impensable sin el manejo de las distribuciones espaciales de las redes viales y la ubicación de los potenciales orígenes y destinos, ya sean estos puertos, almacenes, fábricas, hospitales, estaciones de policía o cuerpos de bomberos, por citar solo una pequeña parte.

Sin embargo, la realización de los análisis espaciales involucrados en todos estos avances impone como requisito indispensable disponer de las coordenadas de los objetos o fenómenos bajo estudio. En efecto, será imposible establecer una correlación entre diversas fuentes de contaminantes químicos, y los casos detectados de una determinada enfermedad respiratoria,

si no se conocen las coordenadas exactas de las residencias de los enfermos y de las potenciales fuentes de contaminación.

De modo que la introducción de las potentes herramientas que ofrecen los GIS, con su enorme valor para los procesos de toma de decisiones, depende directamente de la existencia en las bases de datos de las coordenadas de ubicación de los objetos y fenómenos.

A esta información de carácter espacial, dada por las coordenadas geográficas, se le denomina georreferencia explícita y, en sentido general, no se encuentra disponible en la mayoría de las bases de datos operacionales hoy en día, fundamentalmente las que corresponden a registros tomados en etapas anteriores a la aparición de los sistemas globales de posicionamiento (GPS), que indudablemente han dado un vuelco en la disponibilidad de información espacial.

Históricamente, y aún en nuestros días, la caracterización espacial de los objetos y fenómenos en un importante grupo de actividades, se ha basado en la utilización de las direcciones postales. A ésta se le ha denominado entonces georreferencia implícita y, si bien no deja de constituir una variable de carácter espacial, no sirve a los fines descritos anteriormente, al menos directamente.

La geocodificación es entonces el proceso de determinación de las coordenadas que corresponden a una dirección postal o nombre geográfico dado. Por su parte, la selección de la dirección postal que más se acerca a unas coordenadas geográficas conocidas se denomina geocodificación inversa.

Resulta importante, por tanto, disponer de herramientas computacionales que permitan implementar los procesos de geocodificación en dos variantes:

- un servicio en línea que posibilite ubicar rápidamente sobre un mapa en la pantalla de una computadora el lugar que corresponde a una referencia geográfica textual (una dirección postal o simplemente un nombre geográfico),
- una aplicación informática que, usando el mismo servicio, pueda actuar sobre una base de datos y convertir masivamente todas las direcciones postales contenidas en sus registros en coordenadas geográficas.

Visto así, en su formulación teórica, pudiera considerarse el problema de la geocodificación como ya resuelto. De hecho, existen varios paquetes de software para GIS que ofrecen soluciones para la implementación de servicios de geocodificación (p. ej. MapInfo, ArcGIS y el propio gestor de bases de datos Oracle). Incluso, puede resultar familiar para muchas personas el haber viajado a bordo de taxis en ciudades del mundo desarrollado, en los que se ha instalado un sistema que incluye una pantalla en la que, sobre un mapa, se muestra el destino del viaje, la posición actual del taxi e incluso se dan indicaciones orales al chofer sobre la ruta a seguir.

Sin embargo, la situación actual con respecto a la existencia de servicios de geocodificación en el mundo, no es la que pudiera esperarse de la imperiosa necesidad existente y de estas premisas tecnológicas. Esto se debe a factores como los que se relacionan a continuación:

- Las direcciones postales tienen raíces históricas y culturales, de modo que su estructura

puede variar notablemente de un país a otro. Esto convierte a la geocodificación en un proceso complejo y específico de cada país.

- Por otra parte, el proceso de geocodificación es altamente dependiente de la existencia de una base de datos de referencia que contenga todos los elementos que pueden ser parte de una dirección, con sus coordenadas geográficas correctas. El proceso de construcción y actualización permanente de dicha base de datos, por diversas razones, es altamente complejo y costoso.

Si bien la situación en los países desarrollados no puede considerarse aun ni mucho menos perfecta, en los países en desarrollo (o al menos en la mayoría de ellos) apenas puede hablarse de avance alguno en la temática de la geocodificación, básicamente por la inexistencia de bases de datos de referencia e incluso la ausencia de sistemas formales de direcciones postales en grandes áreas residenciales caracterizadas por la extrema pobreza.

En el presente artículo se exponen las características principales del proceso de implementación y despliegue de un servicio nacional de geocodificación para el enfrentamiento del delito en la República de Cuba, así como los resultados obtenidos de su aplicación a diferentes bases de datos en ámbitos diversos de la actividad humana.

CARACTERÍSTICAS GENERALES DEL PROCESO DE GEOCODIFICACIÓN.

El proceso de geocodificación requiere de tres elementos: estilos y reglas (modelos) de direcciones, la base de datos de referencia (elementos de las direcciones con sus descripciones espaciales) y los algoritmos de procesamiento (Goldberg et. al. 2007).

Cuando se busca una dirección usando un geocodificador, se identifican (usando las reglas) los elementos estructurales y los atributos de la dirección (Figura 1). Estos atributos se buscan entonces en los datos de referencia y se eligen candidatos con atributos similares a los buscados. A cada candidato se le asigna una puntuación indicando la similitud entre lo buscado y lo encontrado y se obtiene, como salida del proceso, los candidatos de más alta puntuación.

Finalmente, para cada candidato se realiza el cálculo de las coordenadas que, según el caso, puede corresponderse directamente con las coordenadas del elemento en la base de datos de referencia (p. ej. cuando la dirección postal representa una esquina de dos calles) o requiere de un proceso de interpolación espacial. En el caso de un segmento de calle, por ejemplo, la estimación se realiza tomando como base los rangos de valores posibles para los inmuebles ubicados en cada acera (los pares y los impares) obteniendo la ubicación aproximada del número postal dado de acuerdo al lugar que ocupa dentro de dichos rangos.

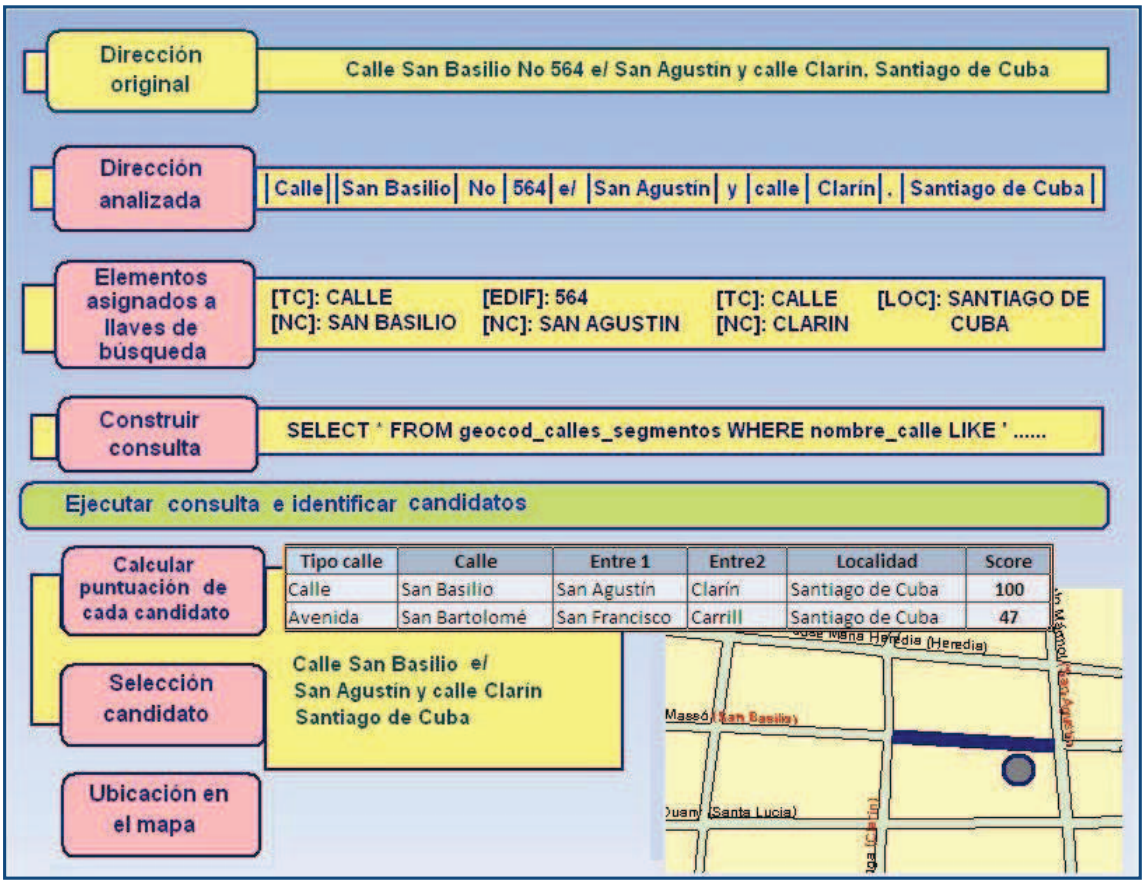

Figura 1. Pasos básicos del proceso de geocodificación.

Los servicios de geocodificación brindan esencialmente tres funcionalidades básicas: geocodificación en línea (dada una dirección se obtienen varios candidatos y el usuario puede escoger el mejor o refinar la búsqueda), geocodificación en lote (dado un conjunto de direcciones, para cada una se escoge al mejor candidato y se obtiene un conjunto de salida), y geocodificación inversa (dadas las coordenadas de un punto se obtiene la dirección que más se ajusta al punto dado).

Para la geocodificación (en lote) de direcciones almacenadas en un fichero o una Base de Datos (BD), se requiere de una aplicación que implemente el proceso de invocación del servicio para un grupo de registros, reciba los resultados de la geocodificación y salve los valores correspondientes.

Dada la enorme importancia que revisten actualmente los procesos de geocodificación, muchas y muy diversas son las herramientas que se ofrecen internacionalmente para la implementación de diferentes soluciones. Así, pueden encontrarse desde páginas Web en las que el usuario puede gratuitamente ubicar una dirección postal sobre un mapa interactivamente hasta juegos de datos en diversos formatos para alimentar motores de geocodificación de otros fabricantes.

Varias revisiones han sido presentadas en la literatura científica con amplias evaluaciones de este tipo de soluciones. En Swift et. al. (2008) aparece una reciente y amplia compilación de alternativas.

Al abordar el despliegue de un servicio de geocodificación de carácter nacional, se requiere básicamente de tres componentes: la definición de los modelos de direcciones utilizados en el país, la selección de la herramienta de software que implemente los algoritmos de geocodificación y la construcción de la Base de Datos de referencia.

En las secciones que siguen se exponen las características principales de estos componentes, los candidatos evaluados, los criterios valorados para su selección, así como las adaptaciones efectuadas para conformar una solución informática coherente, que sirviera de base para el despliegue de un servicio de geocodificación nacional en las condiciones de un país en desarrollo.

MODELOS DE LAS DIRECCIONES CUBANAS.

Las direcciones postales, como sistema de referencia para la localización sobre la superficie terrestre, se han establecido en un prolongado y complejo proceso a lo largo de la historia de la sociedad humana.

Un concepto básico e intuitivo de dirección es: una descripción que incluye nombres y algunas piezas de información complementaria, permitiéndole a las personas identificar unívocamente un lugar. Bajo este concepto se pueden reconocer los siguientes tipos de direcciones:

- Dirección postal. Descripciones estructuradas conteniendo una jerarquía de lugares (ej.: país, provincia, municipio, localidad, calle) y piezas de información complementaria (ej.: número o nombre de inmueble, número de apartamento) usados para identificar un lugar. Los códigos postales constituyen un atajo a zonas o localidades y son elementos redundantes dentro de la dirección.
- Nombre de lugar: Denominación de lugares bien conocidos (puntos de interés), ya sean naturales o hechos por el hombre (edificaciones).
- Dirección relativa: Descripción de la ubicación de un lugar dada por el nombre de otro lugar y alguna información complementaria, como distancia y/o sentido de dirección, para indicar cercanía (ej.: a una cuadra del estadio); o la combinación de dos lugares (ej.: entre la terminal de ómnibus y la plaza de San Carlos).

El uso combinado de estos tipos en cada país ha dado como resultado especificidades, si bien, en sentido general, en el último siglo los procesos universales de globalización y convergencia derivados del comercio, el turismo, las migraciones, etc. han incidido en el establecimiento de modelos básicos, internacionalmente aceptados, para los contextos urbano y rural.

Así, es muy común encontrar en las ciudades el estilo conformado por el nombre de la calle acompañado por el número del inmueble y al final los elementos de la división político administrativa en orden jerárquico ascendente a que corresponde, por ejemplo, municipio – provincia – país.

Adicionalmente, en muchos casos, se utiliza el código de zonificación postal que resulta de gran importancia en la simplificación de los voluminosos procesos de clasificación de cartas y bultos dentro de la actividad de la industria postal. De hecho, se han realizado importantes

esfuerzos en diferentes países para establecer la obligatoriedad en el uso de los códigos postales.

En la temática relativa al sistema de direcciones, Cuba, pequeña isla del Caribe con 11 millones de habitantes y 110 000 kilómetros cuadrados, no escapa a las regularidades descritas, tanto por la existencia de los modelos más comunes, como también por la presencia de importantes especificidades que resulta imposible ignorar en cualquier intento de establecer un servicio nacional de geocodificación. Algunas de las características identificadas son las siguientes:

- Bajo nivel de estandarización de la estructura de las direcciones postales en la sociedad.
- Uso extendido en las direcciones postales de zonas urbanas de las entrecalles, como elemento explícito en el texto de la dirección.
- Existencia en pueblos de zonas rurales de calles que no tienen nombre.
- Muy bajo nivel de utilización de los códigos postales.
- Diversidad de criterios en las formas de numeración postal que puede estar asociada unas veces a las calles y otras a las regiones. Coexistencia de diferentes sistemas de numeración correspondientes a diferentes momentos históricos.
- Diferentes formas de presentación de los prefijos y sufijos para denominar las calles. Ej.: "Ave 5ta Norte" o "5ta Ave Norte".
- Abundancia de nombres alternativos, tanto para las calles como para los asentamientos humanos.
- Amplia utilización en las direcciones de ciudades de nombres asignados a los vecindarios (barrios, repartos, comunidades) que sirven a su vez como elemento de desambiguación.

Con el objetivo de caracterizar con precisión los estilos de direcciones utilizados a lo largo de todo el país, se tomó el conjunto de direcciones postales donde había electores inscritos que fue obtenido para el último proceso de elección de delegados de circunscripción en el país, celebrado en abril de 2010. La participación popular en la revisión y actualización de las *Listas de Electores* las convierte en importante fuente de información sobre los diferentes modelos de direcciones usados en el país y su estructura.

Como resultado de este estudio (Figura 2), y teniendo en cuenta además los criterios acumulados por los especialistas del *Carné de Identidad y Registro de Población*, se conformaron los siguientes modelos:

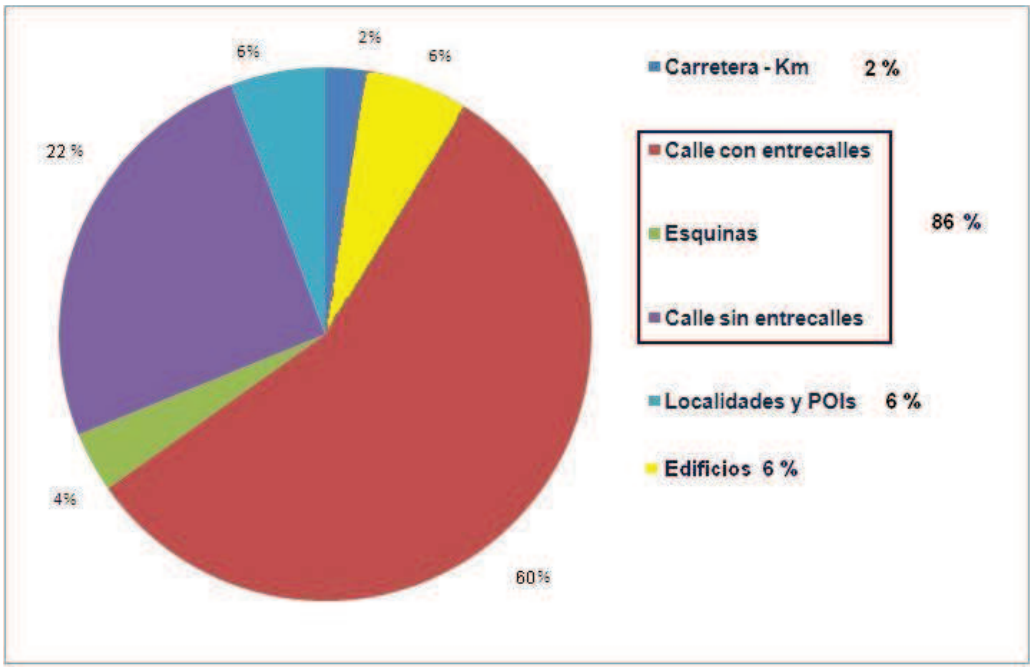

Figura 2. Distribución aproximada del uso de los modelos de direcciones

1. Modelo básico urbano de calles y entrecalles. Este es el utilizado en direcciones como:
   - Monte # 852 Apto. 3 e/ Arroyo y Matadero, Habana Vieja, La Habana, Cuba
   - Ave. 27 No. 4207 e/ 42 y 44, Playa, La Habana, Cuba

   Obsérvese, en ambos casos, la presencia redundante de las entrecalles (precedidas en los textos por la secuencia "e/") que se acentúa en el segundo caso, en el que incluso las propias entrecalles quedan determinadas de forma única por el número del inmueble.

   Este modelo presenta dos variantes: una utilizada para los inmuebles que se encuentran en una intersección vial, como en los ejemplos siguientes:

   - 31 esq. 224, Versalles, Matanzas, Cuba
   - 23 y 12, Plaza de la Revolución, La Habana, Cuba

   O simplemente cuando se omiten las entrecalles como en el modelo tan comúnmente utilizado en otros países.
2. Modelo de referencia lineal, ampliamente utilizado en zonas rurales, como en el ejemplo siguiente:
   - Carretera a Viñales Km 4, José María Pérez, Pinar del Río, Pinar del Río, Cuba
3. Modelo en el que se utiliza como referencia un asentamiento poblado o un punto de interés, como en los ejemplos siguientes:
   - Finca Los Serafines, Sibanicú, Camagüey, Cuba
   - Los Mangos, Amancio, Las Tunas, Cuba
4. Modelo utilizado en un amplio conjunto de urbanizaciones edificadas en las últimas décadas, en las cuales no existe una amplia red vial o no es utilizada en las direcciones.

En este caso se utiliza como referencia solamente el número del edificio y el nombre de la urbanización (o una zona dentro de ésta), por ejemplo:

- Edif. 674 Apto. 30, Alamar 19, Habana del Este, La Habana
- Ed. Q 66 Apto. 6, Micro 7, Distrito José Martí, Santiago de Cuba, Cuba

Basados en estos criterios, se definió un conjunto de reglas para analizar los textos de las direcciones y extraer de ellas los elementos claves para la búsqueda.

Se establecieron entonces dos casos. En el primero, asociado al modelo de edificios, la dirección se inicia por alguno de los prefijos identificados para edificio ("EDIF", "ED", etc) y además no incluye los separadores "E/", "ESQ" ni " Y " que indican la presencia de alguna calle.

En el otro caso, una dirección está compuesta por un elemento que denota calle y después puede estar seguido por un elemento de referencia lineal (debido la presencia de "Km"), o por un elemento que denota segmento (debido a la presencia de "e/"), o por un elemento que denota intersección (debido a la presencia de "y"), o por algún elemento que denote localidades (debido a la presencia de "," o a la presencia de un tipo de localidad como "pueblo").

En este también se incluye el modelo de localidades y puntos de interés. Se presenta cuando el elemento calle no está seguido por nada. Eso implica que si existe la presencia del tipo de calle, entonces de seguro es una calle, pero si no existe dicho elemento, entonces el elemento calle (quedaría solo el nombre) puede ser interpretado como el nombre de un punto de interés o el nombre de una localidad.

CONSTRUCCIÓN DE LA BASE DE DATOS DE REFERENCIA.

Como se ha mencionado antes, un elemento esencial en la geocodificación es la base de datos de referencia. En este caso, la base fue construida a partir de varios conjuntos de datos provenientes de diversos suministradores nacionales.

La fuente principal fue la cartografía digital generada y mantenida por el Grupo Empresarial GeoCuba, proveedor oficial de esta información en el país, pero también se utilizaron algunos registros de la Oficina Nacional de Estadísticas e Información (ONEI), Correos de Cuba y la Oficina Nacional de Identificación y Registros.

La cartografía obtenida de GeoCuba consistió en el resultado de la integración en un solo juego de datos del mapa topográfico 1:100,000, y los callejeros de los asentamientos poblacionales urbanos del país en escalas entre 1:20,000 y 1:5,000.

Este juego de datos se encuentra en formato MapInfo y consta de las capas temáticas siguientes:

- División Político Administrativa conformada a su vez por respectivas capas de Municipios y Provincias.
- Polígonos con las delimitaciones exteriores de todos los asentamientos poblacionales urbanos y rurales según el registro oficial de la Oficina Nacional de Estadísticas.
- Redes viales de las ciudades y otros asentamientos urbanos.

- Red de autopistas y carreteras.

Por otra parte, desde el punto de vista de los requerimientos de la Base de Datos de referencia, la cartografía de GeoCuba presentó un grupo de limitaciones, algunas de las cuales pudieron ser aliviadas o eliminadas en el proceso de construcción de la misma, y otras se convirtieron en limitantes del servicio implementado. Entre otras pueden señalarse:

- Ausencia de la mayoría de los nombres de carreteras y autopistas en las capas correspondientes.
- Los callejeros de las ciudades y pueblos han sido cartografiados para su utilización en mapas impresos fundamentalmente, de modo que presentan dificultades para su empleo en la geocodificación. Un ejemplo claro se encuentra en la presencia de avenidas de varios carriles con separadores, todos los cuales han sido digitalizados y aparecen como objetos diferentes. Otra situación de este tipo aparece en la inclusión en la cartografía de gran cantidad de accesos a centros socioeconómicos como parte de la red vial pero que no forman parte de las direcciones postales.
- Errores de tipografía en los nombres de los objetos y ausencia de criterios de normalización para su empleo.
- Existencia de calles sin división en segmentos.

Partiendo de estos datos, y tras un proceso de Extracción-Transformación-Limpieza (ETL) se creó la base de datos de referencia sobre el gestor de Bases de Datos Oracle (utilizando su extensión espacial Oracle Spatial), incluyendo las capas correspondientes a los elementos estructurales necesarios para la geocodificación. Algunas de las tareas principales acometidas en este proceso fueron desarrolladas como procedimientos automáticos programados en el lenguaje de gestión de la base de datos y se enumeran a continuación:

- Procesamiento inicial de la cartografía. Incluyó básicamente la revisión ortográfica de los nombres y la normalización de los estilos utilizados para calles, asentamientos humanos y otros elementos presentes en las direcciones postales.
- Digitalización de repartos y barrios en las ciudades principales.
- División automática de segmentos múltiples y composición automática de segmentos fragmentados. Dadas las características de la cartografía de Cuba, explicadas anteriormente, no siempre los segmentos de calles estaban representados por uno y solo un objeto. (Figura 5).

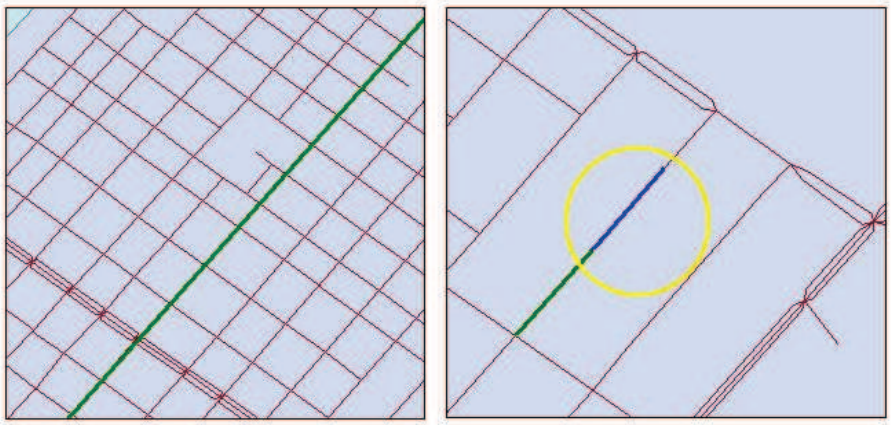

Figura 5. Segmentos de calle múltiples y fragmentados.

- Adición automática de algunos alternativos, básicamente los asociados al uso indistinto de ordinales y cardinales para los nombres de calles con valores numéricos, por ejemplo, "1", "1ra", "Primera" e incluso "Uno".
- Construcción automática de las calles a partir de los segmentos.
- Determinación automática de las intersecciones entre las calles construidas.
- Determinación automática de las entrecalles de cada segmento.
- Detección automática de situaciones de "T". Se le ha denominado de esta forma a la topología resultante cuando un segmento de calle no atraviesa completamente la otra calle en una esquina. (Figura 6). Cuando esto ocurre los inmuebles de una acera tendrán direcciones diferentes a los de la otra en una misma calle.

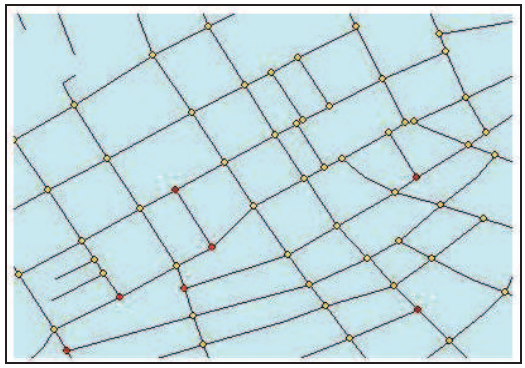

Figura 6. Situaciones de "T".

- Duplicación automática de objetos (segmentos de calle e intersecciones) en áreas colindantes a las fronteras de municipios. (Figura 7). La capital del país, Ciudad de La Habana, a diferencia del resto de las provincias, en que las ciudades cabeceras ocupan un solo municipio, está formada por varios municipios, de modo que las fronteras que los separan están dadas en muchos tramos por ejes viales con una acera en un municipio y la otra en un municipio diferente. Como consecuencia, en los segmentos de calle aledaños a estas fronteras se presenta cierto nivel de confusión con respecto al municipio en que realmente se encuentra determinado inmueble, con lo cual las direcciones postales correspondientes pueden darse en uno de los municipios o en el otro.

Figura 7. Zona de confluencia de tres municipios.

- Obtención (aún en proceso) de la ubicación de edificios que forman parte de urbanizaciones en las que la dirección postal está dada solamente por el vecindario y el número del edificio. Se seleccionaron las urbanizaciones que agrupan en el país la mayor cantidad de direcciones postales con este modelo y se llevó a cabo un proceso combinado de gabinete para identificar los edificios en imágenes satelitales (Figura 8) con el trabajo de campo para el levantamiento de los números postales correspondientes.

Figura 8. Área residencial de Alamar, al este de La Habana, y donde las direcciones postales están dadas por el número del edificio y el nombre de la zona.

De esta forma, los conjuntos de datos originales sirvieron de base para conformar los elementos de referencia presentes en los diferentes modelos de direcciones utilizados en el país y que son:

- División Político Administrativa conformada por los niveles de provincia y municipio (14 y 169 objetos respectivamente).
- Asentamientos humanos urbanos y rurales según el registro oficial de la ONEI conciliado con la cartografía de GeoCuba (cerca de 7,000 objetos).
- Repartos, barrios y otras comunidades identificadas dentro de las ciudades cabeceras provinciales y algunos otros núcleos poblacionales importantes (alrededor de 580 objetos).
- Segmentos de calles, carreteras y autopistas en ciudades y otros puntos poblados provenientes de los callejeros de GeoCuba en las escalas desde 1:5,000 hasta 1:2,000 (más de 195,000 objetos).
- Intersecciones de calles (cerca de 106,000 objetos).
- Diagramas georreferenciados a escalas aproximadas de 1:5,000 de edificios de viviendas en conjuntos poblacionales en los que las direcciones postales se basan en referencias a los propios edificios (actualmente cerca de 1,200 objetos).
- Puntos de interés conformados a partir del registro nacional de nombres geográficos y otras fuentes públicas de entidades políticas, socioeconómicas y culturales (más de 60,000 objetos).

RESULTADOS.

El servicio de geocodificación ha sido desplegado dentro de la infraestructura de datos espaciales (IDE) para el enfrentamiento del delito y se encuentra operacional desde septiembre 2010. A partir de su puesta en marcha ha sido utilizado ampliamente en sus dos modalidades, es decir, como localizador en los visores de mapas que ofrece la IDE, y como apoyo a aplicaciones que incluyen opciones de representación de información sobre mapas y que son desarrolladas por diferentes equipos dentro de la institución.

También se ha utilizado la herramienta de geocodificación en lote en Bases de Datos Oracle para procesar diferentes juegos de datos, entre otros los de accidentes de tránsito, y de algunos tipos de delitos que requieren atención especial.

En el caso del servicio de geocodificación en línea, resulta difícil establecer con precisión métricas objetivas de calidad de la búsqueda, debido principalmente a la riqueza de posibilidades que ofrece la interacción del usuario con la herramienta que permite acomodar los pedidos a la información de referencia existente.

Por otra parte, en el caso de las bases de datos históricas que incluyen direcciones postales, es donde realmente se pone de manifiesto la potencia de los algoritmos de reconocimiento y búsqueda para enfrentar una amplia gama de situaciones, patrones y criterios de redacción diversos en la conformación de las direcciones postales ante la inexistencia ya mencionada de normas para esta actividad.

Se ha tomado entonces, como criterio para caracterizar la calidad del servicio desplegado, los resultados obtenidos en la realización de un conjunto de procesos de geocodificación en lote sobre bases de datos de diversa procedencia.

Para esto se han tomado los siguientes juegos de datos, en los cuales no se realizó ningún pre procesamiento para limpiar o normalizar las direcciones:

- Muestra aleatoria de 10,000 direcciones de viviendas del registro de identidad nacional distribuidas por provincias en proporción a la población de cada una.
- Lista de las principales oficinas postales del país (998 registros).
- Muestra aleatoria con más de 2,000 entidades económicas, sociales y culturales de la provincia de Ciudad de La Habana.
- Registro de los teléfonos públicos de la provincia de Ciudad de La Habana (alrededor de 14,000 direcciones).

En el proceso de geocodificación de estas direcciones se obtuvo, aproximadamente, el 50% de casos exitosos de forma global, aunque una característica importante de las estadísticas obtenidas fue la elevada variabilidad por territorios (entre provincias, entre zonas urbanas y rurales), como era de esperar de las diferencias socioeconómicas y culturales existentes (Tabla 1).

Juego de datos	Muestra	Encontradas	%
Carné de identidad	10000	4834	48.34
Oficinas postales	998	435	43.59
Entidades Ciudad Habana	2093	938	44.82
Teléfonos públicos Ciudad Habana	13925	7303	52.45

Tabla 1. Resultados en la geocodificación de varios juegos de datos.

Otro aspecto derivado de estas valoraciones preliminares, fueron las notables diferencias existentes entre los resultados en juegos de datos de procedencia diferente, lo cual confirmó la necesidad de realizar el pre procesamiento de los juegos de datos con vistas a obtener niveles de efectividad más cercanos a los máximos implícitos en el uso de la herramienta.

En este primer intento de obtener estimados de la calidad del servicio, se incluyó también una comparación con servicios internacionales de diseño similar al implementado. Así, se procesó la muestra obtenida del carné de identidad utilizando el servicio que ofrece Google para la geocodificación en línea (Gilmore 2006a, b).

Los resultados en este caso fueron inferiores a los nuestros en todas las provincias (Figura 9), pero especialmente en un grupo en el que apenas superó el 10 %. Obsérvese que las barras de la derecha en cada provincia representan los resultados obtenidos con el servicio de Google y las de la izquierda corresponden al servicio presentado en este artículo.

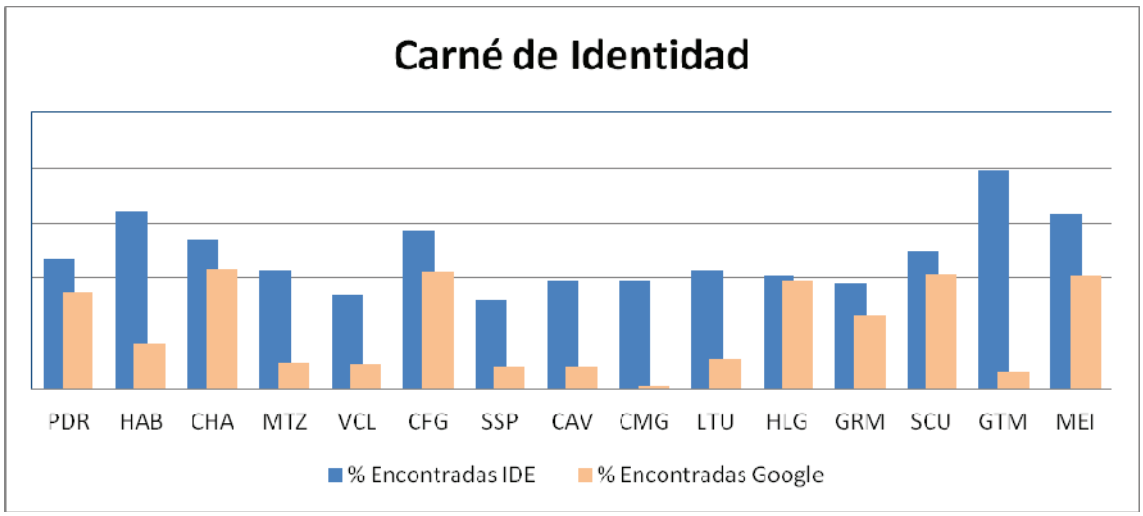

Figura 9. Comparación con el servicio de geocodificación de Google.

Finalmente, se ha desarrollado un proceso de evaluación cuantitativa y cualitativa de la calidad del servicio tomando como base el Registro de Electores a nivel nacional que incluye, como se ha dicho, más de 2 millones 700 mil direcciones de todo el país. Este registro ha sido objeto de un trabajo de procesamiento previo en que las direcciones fueron limpiadas, normalizadas y estructuradas en campos diferentes para cada uno de los elementos correspondientes a cada modelo.

A continuación se presentan los resultados de este estudio que muestran el nivel alcanzado distribuido por provincias, y que ofrece el servicio actualmente.

Provincia	Total	Encontradas	No encontradas	%
Pinar del Río	124,796	85,579	39,217	68.58
La Habana	195,469	168,064	27,405	85.98
Ciudad de La Habana	648,907	458,166	190,741	70.61
Matanzas	189,071	97,113	87,738	51.36
Villa Clara	222,194	121,483	100,711	54.67
Cienfuegos	91,655	56,785	34,870	61.96
Sancti Spíritus	97,087	52,891	44,196	54.48
Ciego de Ávila	103,961	68,388	35,573	65.78
Camagüey	215,770	111,165	104,605	51.52
Las Tunas	110,496	80,294	30,202	72.67
Holguín	199,368	125,957	73,411	63.18
Granma	153,029	76,640	76,389	50.08
Santiago de Cuba	227,820	124,259	103,561	54.54
Guantánamo	96,712	62,693	34,019	64.82
Isla de la Juventud	27,022	14,203	12,819	52.56
Totales	**2,703,357**	**1,703,680**	**995,457**	**63.02**

Tabla 2. Resultados por provincias en la geocodificación del Registro de Electores.

En términos de velocidad de respuesta, el uso del servicio para la geocodificación masiva mantiene niveles aceptables. Se realizó un estudio de la relación de la velocidad de procesamiento con respecto al tamaño de los lotes, obteniéndose, para las condiciones del

despliegue, los valores que pudieran ser más convenientes (Figura 10). En el caso de la geocodificación en línea, el nivel de respuesta promedio del servicio se mantiene alrededor de los 6 segundos para la búsqueda de una dirección.

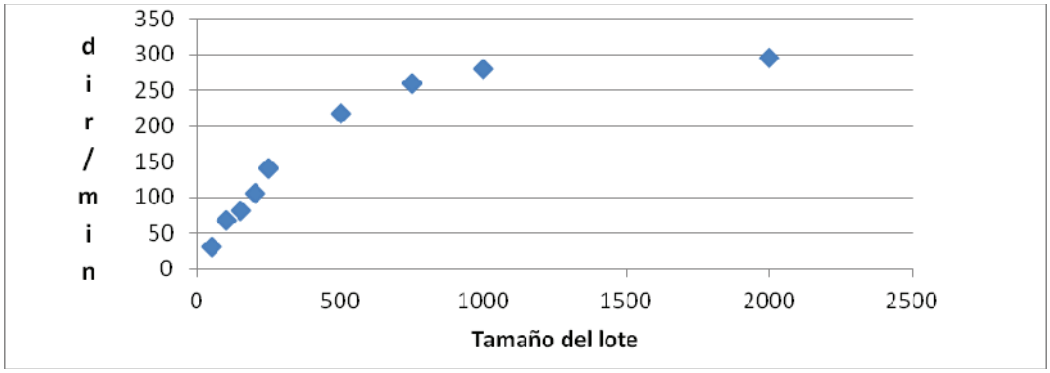

Figura 10. Relación de la velocidad de geocodificación con respecto al tamaño del lote.

CONCLUSIONES

En este trabajo se han presentado los resultados más importantes del proceso de implementación y despliegue de un servicio de geocodificación en la Infraestructura de Datos Espaciales para el enfrentamiento del delito de la República de Cuba.

Como ha quedado demostrado, el servicio desplegado ha permitido georreferenciar, de forma automática, juegos de datos con información de temáticas diversas lo que a su vez ha posibilitado realizar análisis espaciales de gran utilidad en el enfrentamiento del delito. Esto puede ser extendido a otras áreas como la salud, el transporte, etc.

Especialmente, el procesamiento del Registro de Electores a nivel nacional ha sido una experiencia de alto valor, tanto para profundizar en el conocimiento de las características de las direcciones cubanas, como en el perfeccionamiento de la base de datos de referencia.

De esta forma, además, el país cuenta con una herramienta de elevado valor para su amplio uso en diversas aplicaciones en que sea necesaria la localización de objetos y fenómenos a partir de sus direcciones postales.

El trabajo, sin embargo, representa solo el comienzo. Como se ha podido apreciar, los niveles de efectividad del servicio son aún bajos (Jacquez and Romel, 2009), pero se conocen claramente las áreas en las cuales es necesario continuar trabajando para elevar progresivamente los indicadores de calidad del servicio. Por otra parte, en este mismo sentido de la calidad, resulta impostergable establecer métricas cuantitativas y objetivas que permitan establecer con precisión los niveles de calidad del servicio y los avances que se alcancen (Davis and Fonseca, 2007).

Otras tareas de desarrollo futuro identificadas son las siguientes:

- Completar el trabajo de digitalización de edificios en urbanizaciones cuyas direcciones están dadas como referencia a los propios edificios y la localidad.
- Completar el trabajo de digitalización de los repartos, barios y otras zonas residenciales definidas dentro de las ciudades principales.

- Establecer un programa de trabajo conjunto con las entidades de cartografía y transporte del país para completar el registro (cartografiado) de las carreteras del país, incluyendo las marcas de distancia.
- Completar las listas de números postales existentes en cada calle con vistas a habilitar la interpolación en los segmentos y aumentar en nivel de exactitud de las coordenadas obtenidas en el modelo de direcciones más utilizado en el país.
- Formalizar un estudio estadístico completo con muestras significativas que permita establecer los niveles de exactitud real con criterios científicamente fundamentados en diferentes zonas y contextos del país, y compararlos con referencias internacionales (Jacquez and Romel 2009).

REFERENCIAS BIBLIOGRAFICAS

Crossier, Scott, 2004, ArcGIS 9. Geocoding in ArcGIS. (Redlands: Enviromental Systems Research Institute).

Davis, Clodoveu A, and Frederico T. Fonseca, 2007, Assessing the Certainty of Locations Produced by an Address Geocoding System. Geonformatica, 11:103-129.

Enviromental Systems Research Institute (ESRI), 2009, ArcGIS 9.3 Geocoding Technology. (Redlands: ESRI).

Enviromental Systems Research Institute (ESRI) , 2010. ArcGIS Desktop 9.3 Help. Commonly used address locator styles. (Redlands: ESRI).

Gilmore, W. J., June 2006, Introducing Google Geocoding Service. The Network for Technology Professionals. http://www.developer.com/tech/article.php/3615681/introducing-google-geocoding-service.htm.

Gilmore, W. J., July 2006. Performing HTTP Geocoding with the Google Maps API. The Network for Technology Professionals. http://www.developer.com/db/article.php/3621981/Performing-HTTP-Geocoding-with-the-Google-Maps-API.htm

Goldberg, Daniel W., Wilson John P., and Craig A. Knoblock, 2007, From Text to Geographic Coordinates: The Current State of Geocoding. Journal of the Urban and Regional Information Systems Association. 19 (1): 33-46.

Jacquez, Geoffrey M, and Robert Romel, 2009, Local Indicators of Geocoding Accuracy (LIGA): Theory and Application. International Journal of Health Geographics 8 (1): 60.

Kothuri, Ravi, Albert Godfrind, and Euro Beinat, 2007, Pro Oracle Spatial for Oracle Database 11g. (New York: Apress).

Murray, Chuck, 2008, Oracle 11g Spatial Developer Guide. (Redwood: Oracle Press).

Shea, Cathy, 2007, Oracle 11g Text Reference. (Redwood: Oracle Press).

Swift, Jennifer N., Daniel W. Goldberg, and John P. Wilson, 2008, Geocoding Best Practices: Review of Eight Commonly Used Geocoding Sustems. Technical Report No. 10, GIS Research Laboratory, University of Southern California (Los Angeles, CA).

Tang, Agatha, and Kristin Clark, 2003, ArcGIS 9. Geocoding Rule Base Developer Guide. (Redlands: ESRI).